Comparability and Examination Performance: Technical and Social Approaches to Its Study

Thesis

UNRESTRICTED

COMPARABILITY AND EXAMINATION PERFORMANCE: TECHNICAL AND SOCIAL APPROACHES TO ITS STUDY

ANN CHRISTINE BENSON B.Sc. M.Ed.

THE OPEN UNIVERSITY

DOCTOR OF PHILOSOPHY

EDUCATION

15 DECEMBER 2008

Submission date: 30 June 2008
Date of award: 26 Jan. 2009

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

ABSTRACT

This thesis is concerned with examination comparability and the assumption that achieved grades in GCSE examinations have common currency across subjects.

Technical treatments commonly used to investigate examination comparability are discussed along with the assumptions upon which they are based and their limitations. A variety of technical treatments, taking into account population sampling, tiering, coursework and cognitive skill demands, are used to investigate comparability for GCSE science results from Welsh and English examining groups. Examination comparability is shown to be undermined by fluctuations in relative 'difficulty' across time, different correlations between subjects, curriculum changes, and sub-group effects.

Interviews with science teachers are then related to the technical findings to examine schools', departments' and individuals' responses to national assessment structures and practices and how these mediate 'gradeness'. The interviews' initial focus on teachers' tier entry decisions reveals that their judgements about students are constituted through interaction between their beliefs about mind, subjects and gendered behaviours amongst others, departmental and school practices, and wider social influences to do with national assessment and examining group policies and practices.

The interviews show how structures and beliefs shape arena practices and teachers' practice, the consequences for students' access to science, and the consequent validity of assessments of their science 'achievements'. The allocation of students to ability groups as they enter secondary school and the interactions between these groups and KS3 SAT tier allocation effectively 'lock' students on to an assessment pathway from Year 7 – a pathway which school structures make it almost impossible to break away from. The findings show how school practices and individual practice can disrupt or compound this and the consequences for students' access to learning opportunities, which, the thesis argues, is a major source of invalidity in assessment that comparability studies cannot begin to take account of.

ACKNOWLEDGEMENT

Special tribute is paid to my supervisor, Professor Patricia Murphy, for her patience, encouragement and constructive criticism in guiding me through the writing of this thesis. I am grateful for her unfailing support.

CONTENTS

LIST OF FIGURES

LIST of TABLES

CHAPTER 1

Introduction: the genesis and aims of this research

1.1 Initial concerns: students' differential science examination performance

This research study is concerned with examination comparability and the assumption that achieved grades in national examinations have common currency across subjects. I first became interested in the notion of 'gradeness' being stable across subjects, and challenges to it, from my teaching and examining experiences in the 1990s.

I taught chemistry to 11 – 18 year old students from 1968 until 1993 when I became a university lecturer with teaching responsibilities on science Postgraduate Certificate in Education (PGCE) courses. At that time the national General Certificate of Secondary Education (GCSE) was (and continues to be in 2008) taken by students 15 -16 years of age in England, Wales and Northern Ireland. As part of my role as an examiner of GCSE combined science for the Welsh Joint Education Committee (WJEC) during a four year period (1990-93 inclusive), I marked annually some four to five hundred scripts and noted a difference in students' performance across the separate science subjects. Marks gained by students on the chemistry sections of examination papers were proportionally lower than marks gained on either the separate biology or physics sections of the same examination papers. This differential performance was confirmed by the WJEC Chief Examiner's reports of the same Single and Double Award GCSE combined science examinations. WJEC Chief Examiner's reports for each of the 1989 to 1992 examinations also contained comments regarding the disparity in performance of students on the biology, chemistry and physics questions.

> *For many years, the chemistry parts of the science papers have either been ignored or very badly answered by the majority of the candidates* [students taking the GCSE]. *Last year the quality of the answers to the chemistry sections of the papers showed a slight improvement; alas this year the situation has again worsened. Candidates were unable to cope with even the simple aspects of chemistry, consequently almost all parts relating to chemistry, even question 1, became differentiating elements.*
>
> (Lapham, 1989)
>
> *It is very disappointing to report once again, that the standard of achievement in chemistry still gives great cause for concern. Many of the very easy concepts and facts are unknown*

by many, many candidates.

(Lapham, 1992)

The Chief Examiner's reports did not provide statistical evidence to substantiate these statements. However, descriptions of students' answers to individual questions were included in the reports. These descriptions were examples of inadequate responses that were judged to be typical. The descriptions demonstrated that students were having greater difficulty in achieving high scores in the chemistry questions than either the biology or physics questions.

During public meetings at this time (Caerleon Teachers' Convention, Gwent, 1980s), the Secretary of the WJEC referred to comparability studies carried out in the 1960s (Robins, 1972), claiming that they revealed chemistry to be a 'harder' subject than either physics or biology in WJEC GCE 'O' level examinations. The implication was that students taking a range of GCE 'O' level subjects would find it more difficult to achieve a high grade in chemistry than in most other subjects. Neither statistical evidence nor underlying reasons for this apparent disparity were offered during these presentations. However, during subsequent discussions between myself and the then Assistant Secretary for research and development at WJEC (WJEC, 1995, personal communication), and after the introduction of GCSE, statistical evidence was claimed to exist which showed a prevailing phenomenon for both chemistry and Latin at GCSE to be regarded as 'hard' subjects. The Assistant Secretary defined a 'hard' subject in terms of a subject in which there is a tendency for students to obtain a lower grade than in their other GCSE subjects.

Thus, there appeared to be concerns regarding students' performances in chemistry relative to biology and physics in a variety of GCSE examinations offered by WJEC. The same type of concern was being expressed for:

- combined science examinations called GCSE Double Award Science, traditionally taken by students for whom separate science GCSEs is considered to be inappropriate in terms of intellectual demands;
- biology, chemistry and physics as separate science examinations called GCSE Triple Award Science, traditionally taken by students who are considered to be intellectually very able when compared with their contemporaries.

These concerns were expressed by senior examination group staff and chief examiners over several years. The concerns could not be dismissed as being due to particular chief examiners' practices as all GCSE examinations were subject to the same systems from paper construction to grade appeals.

As a science and chemistry teacher in a variety of secondary schools for 25 years, preparing students for WJEC science examinations amongst others, I had mixed reactions to these expressed concerns. From my own teaching experiences I had no evidence to indicate that my students found chemistry to be a 'harder' subject than either physics or biology. Conversely, I had first hand confirmation of students' relative difficulty with chemistry from my own WJEC examining activities. To explore these concerns further, I undertook some informal discussions with science educationists who might have access to evidence regarding students' relative performances on these subjects within Wales.

The Director of Techniquest, Cardiff, and ex science education lecturer at the University of Wales College of Cardiff with extensive contact with science teachers throughout Wales, proved to be the most useful in this respect. Conversations (Techniquest, 1995) with the Director revealed no statistical evidence to suggest that there was any aforementioned disparity. However, he did claim that as a result of his conversations with science teachers throughout Wales in recent years (no dates specified) he had considerable anecdotal evidence of two kinds. First, he reported science teachers' concerns regarding differences in the cognitive skill requirements of the WJEC GCSE separate science examination papers. In general the introduction of GCSE was perceived to have resulted in the production of biology examination papers that showed a proportionally greater increase in cognitive skill demand than those of either chemistry or physics. The cohort of students following the National Curriculum for all of their secondary schooling sat GCSE examinations for the first time in 1995. At this point I was a PGCE science lecturer and this prompted me to involve my PGCE science students at the University of Oxford in an analysis of the 1993, 1994 and 1995 WJEC Triple Award GCSE Science examination papers. At this time WJEC specified the relative percentages of different types of cognitive skills, for example recall, application and evaluation, on the science papers in their syllabuses. This hierarchical representation of knowledge demands and the assumption that item (examination question) demands are stable across students reflects a particular view of knowledge that was enshrined in the science syllabuses of WJEC and in other

GCSE examination groups' science syllabuses at that time. In this view labeling item demands is a device used to ensure balance in score weightings. Both examiners and teachers referred to the representation of the different cognitive demands within and between science subjects as a potential source of incomparability. Using my PGCE students' judgements I determined the percentage of each examination paper's total marks allocated to specific types of cognitive skills (Benson, 1995). The analysis showed a significant change in mark weightings for different types of cognitive skills from 1994 to 1995 for all three of the Triple Award GCSE separate science subjects and in particular for biology. The anecdotal evidence of teachers' concerns regarding WJEC separate science examinations papers appeared to be corroborated by the analysis. Second, the Director's anecdotal evidence in contrast to the Chief Examiner's reports also revealed that science teachers generally perceived their students to be attaining lower grades in biology than in either chemistry or physics when all three GCSE subjects were examined by WJEC. As I had corroborated the anecdotal evidence of teachers' concerns regarding examination paper cognitive skill demands, I was mindful of not dismissing this second concern. Overall, the issue of comparability of students' performances in different science subjects at GCSE was clearly a concern for science teachers, chief examiners and examining group personnel at WJEC, and consequently became the focus of my research.

1.2 Subject comparability in examinations

In general, comparability, at the time of my study, had not been taken to mean that an *individual* taking examinations in different subjects would necessarily attain the same grade in those subjects (Goldstein, 1986). It is generally accepted that individuals develop understanding and skills differentially across subject domains: they may also reasonably be expected to respond differently to different examination papers sat over a period of time, as is the case in GCSE administrative arrangements.

Comparability applied to national examination results is concerned with *groups* of candidates [students taking an examination]. The Schools Council, which had general responsibility for British national examinations until the mid-1980s, set up a Forum on Comparability in the late 1970s. In its associated report, albeit concerning comparability of students' attained grades between different examining groups, it said:

> ... *the expectation is that had a group of examinees followed another board's* [examining group's] *syllabus and taken its examination, they might reasonably be expected to have obtained the same average grade.*
>
> (Schools Council, 1979)

Certainly this was the case during the 1960s and 1970s when several research studies (see Schools Council, 1971 and Schools Council, 1979 as examples) focused on comparing the grades awarded for subjects by different examining groups or on item (question) banking. The Council's published work neither focused on comparing performances of the *same* group of students nor across biology, chemistry and physics examinations (currently in 2008, no comparability study of the same type of group of students' performances in GCSE biology, chemistry and physics examinations over time lies in the public domain). The reports paid no attention to the possibility of other variable effects within examining groups related to students and their experiences or between examination groups and their practices and instruments.

During the 1980s and early 1990s examination groups were reluctant to engage in comparability studies of science subjects. Cresswell (1997) writes extensively about the reasons for this lack of research activity throughout the 1980s in his unpublished thesis. According to Newbold (1995, personal correspondence) the lack of research activity is largely because many examination centres, which are organizations, usually schools or colleges, at which students are prepared for and take an examining group's examinations, split their science entries between examining groups. For example an examination centre might use the Midland Examining Group for Salters chemistry and Northern Examinations and Assessment Board for physics. According to Newbold (ibid.) this results in data that is incomplete and unrepresentative of the national position. In this respect, examination centres in Wales are unique. All state secondary schools in Wales are registered centres for WJEC examinations. Consequently, the vast majority of young people in Wales are prepared for GCSE examinations administered by one examination group, WJEC. The number of students sitting WJEC science examination subjects outside Wales is also relatively small when compared with the home entry. WJEC candidates' performances in science subjects at GCSE may as a whole be viewed as one of the products of educational institutions within Wales.

Consequently, it could be argued that if there are inherent, significant differences in the same students' performances in biology, chemistry and physics, then that disparity should be more likely to manifest itself in a context such as Wales and WJEC, where there is a tradition for students to take all of their examinations with the one examining group, than in England with its greater flexibility of candidature (collective noun for all of the students taking a particular examination). Therefore, I felt that a comparability study of the same group of students' performances in science GCSE examinations in the context of Wales and its associated examination group, WJEC, would have the potential to extend understanding in the field more generally.

1.3 Other influences on the research.

There were other factors at this time (Spring, 1995) that influenced my views of the value of such a comparability study.

Inter-group statistical reports of GCSE (UK) science examinations (Inter-Group Research Committee, 1993) recorded a continuing decline in the numbers of students taking biology, chemistry and physics as separate sciences at GCSE. There had been a 30% fall in the total population of 18 year-olds since 1983 and an even larger percentage decline in the number of students continuing with science studies after the age of 16 and entering for 'A' level biology, chemistry or physics (TES, 25 March, 1994 p. 10). Woolnough (1991) amongst others argued that such conditions disadvantaged the UK compared with other countries in respect of the proportion of the population suitably educated to become qualified engineers.

Professor Alan Smithers, at the time Director of the Centre for Education and Employment Research at Manchester University, argued that there was a link between the unattractiveness of the sciences in our culture and the low number of scientists entering teaching (TES, 25 March, 1994 p. 10). Particularly worrying, said Professor Smithers (ibid.), was the continuing trend for mathematics, physics and chemistry trainee teachers to have poor degrees, with more than a third having a third class honours degree, compared with only about one in twenty of those training to become history or English teachers. This concern was heightened by the publication of statistics (DES, 1993) revealing that up to one third of teachers then teaching chemistry in secondary schools were eligible for retirement within five years.

Arguably, the 1988 Education Reform Act (ERA) heralded the most dramatic changes in the provision of education for children in England and Wales this century. For the first time schools were obliged by law to provide a balanced and broadly based curriculum known as the National Curriculum. Science was (and continues to be in 2008) defined as a core subject and encompassed elements of biology, chemistry, physics and Earth science. This 'balanced' view of science meant that all students, boys and girls, of all abilities, were required to study all of these different science disciplines until the age of 16. As a result of these statutory curriculum changes, there were two major consequences of relevance to this research, though to differing extents:

- Earth science was included as a substantial part of the chemistry science curriculum with a consequent reduction in the chemistry content that could be taught to students aged 11 to 14 and redefined chemistry as taught in schools (Benson, A., 1993). The revised National Curriculum of 1995 to an extent redressed this situation with a reduction in Earth science and an increase in chemistry content, although concern continues to be expressed about students' declining chemical ability from comparative studies across time (Assessment Subject Group, RSC, 2003);
- there was an increase in the amount of science taught and examined by age 16, in the disciplines of biology, chemistry and physics. It was expected that a minority of students would take Single Award science (1 grade point), the majority would take Double Award science (2 grade points) and the most able would take the three separate sciences as Triple Award science (3 grade points).

My deliberations about whether or not to conduct research on GCSE science examinations took place during 1994 and 1995. At this time GCSE was still a relatively 'new' examination, having been first examined in 1988. Whether or not the GCSE was achieving its objectives was a matter of controversy at the time of my deliberations. Three main objectives had been associated with the introduction of the GCSE (NAS/UWT, 1989):

(i) the unification of the dual examination system of General Certificate of Education Ordinary Level (GCE 'O' level) and the Certificate of Secondary Education (CSE), and its replacement with a single set of examinations available to all but a minority of students. This objective was driven by dissatisfaction amongst teachers, parents, employers and educationists with a system which placed a good deal of pressure upon teachers to categorize children during the early years of secondary

education as 'O' level and / or CSE students who were then required to follow different types of syllabuses (ibid.). Governmental desire to raise educational standards by means of assessment played a significant role in unifying the GCE 'O' level and CSE systems;

(ii) the promotion of more practical and investigative methods of learning and assessment (ibid.). In science the GCSE was designed to test experimental work carried out by students in the laboratory over time through the coursework component of the GCSE assessment process. This was largely based upon teacher designed and assessed materials. GCSE science coursework was (and still is in 2008) subject to moderation by examining groups;

(iii) the replacement of student-referenced achievement based on traditional norm-referencing systems of GCE and CSE with a detailed method of criterion-referencing of levels of attainment and grades to enable the development of differentiated papers, which assessed what students could do in relation to the domain as opposed to each other (ibid.)

Sub-effects within any observed differential subject performance were beginning to interest examination researchers at this time (January, 1994). Multiple choice questions had been found by Murphy (1982) to preferentially favour the attainment of boys but these were not used in WJEC's Triple Award GCSE science examinations. The methods of assessment used in the GCSE, particularly coursework, were thought to favour girls, although at this time the findings of Stobart *et al.* (1992) challenged this belief for GCSE English and mathematics. Other sub-effects identified included for example girls' relatively poorer performance than boys in physics on the national monitoring tests of performance conducted by the Assessment of Performance Unit for the Department of Education (APU, 1985a) and differential gendered performances on GCSE examinations in English and mathematics (Stobart *et al.*, 1992). This suggested that any exploration of examination comparability should pay attention to the potential for sex group sub-effects.

The move away from norm-referencing towards broader definitions of achievement defined in terms of specific levels of attainment calls for the formulation of grade criteria which specify in some detail, and in subject-specific terms, the levels of attainment necessary for the award of each GCSE grade (Orr and Nuttall, 1983). Attempts to generate grade criteria for a system of national examining had proved problematic. Currently (2008), such grade criteria have still not been developed.

So it was against a background of:

- declining numbers of GCSE biology, chemistry and physics students;
- declining numbers of students carrying on to study science, and in particular chemistry and physics, at 'A' level;
- declining numbers of well qualified science teachers;
- controversial, major national changes in science education and its assessment;
- controversy as to whether the GCSE had fulfilled its objectives and whether it would ever reach its goal of becoming criterion-referenced with grade criteria,

that my attention was drawn to the apparent disparities in students' performances in the different science subjects at GCSE in Wales. I concluded that a comparability study of students' performances in GCSE separate science subjects would prove useful in adding to the science education and assessment debate within the UK at this particular time by providing insights into GCSE grade awarding policy and practice and recommendations for their change.

1.4 The aims of the research and my learning pathway

I came to this research as a physical scientist trained in quantitative methods of inquiry with ontological and epistemological positions aligned respectively to realism and positivism. The technical nature of previous comparability studies attracted me to this field. In looking at the issue of 'gradeness' my first exploration was of the existing comparability studies all of which treated assessment as a technical rather than a social phenomenon. Comparability in these studies assumed that 'gradeness' was stable across subjects and was investigated using a range of technical means and the quantitative outcomes of examinations. The purposes behind the studies were either to identify the most appropriate technical methods of demonstrating the stability of 'gradeness', or by using technical analysis, to challenge the possibility of the existence of such grade stability across subjects. From my, then, social constructivist perspective my purpose was to understand better what 'gradeness' meant for a particular cohort over a specific time span of educational change and in so doing consider what might influence grade meanings and raise questions about their validity. I drew on quantitative data because grade distributions are the phenomena at the heart of the claims I wished to consider.

Therefore when I came first to this research it was within a quantitative paradigm where I explored how grade stability would play out within particular datasets and how the assumption of grade stability might be challenged using the technical means identified in the literature. My intention was first to explore performance across science subjects and possible sources of performance influences as a means of understanding users' claims and beliefs about differences in examination difficulty and examining group claims about comparability. Alongside this I was also interested to consider the potential mediation of gender as the dropout of girls relative to boys from science prior to the National Curriculum was well established and the potential for sex sub-effects such as those discussed earlier was clear. This issue was not, however, raised by teachers or examiners themselves. I thus first explored:

- students' GCSE separate science subject performances for relationships and any sex group effects (research aim 1);
- whether there were relationships between students' GCSE performances in the different science subjects and variables such as examination paper construction factors (research aim 2).

My overarching interest in the thesis was to raise questions about potential sources of invalidity in assessment claims about grade comparability. The inclusion of investigations into sex group effects would serve to illuminate aspects of this. However, as I progressed with my reading and research my theoretical position underpinning my understanding of assessment was shifting. My research was at the same time disrupted by ill health and this coincided with my increasing interest in viewing assessment as a social force that shapes and is shaped by teachers' practice. Thus I felt an uncomfortable dilemma about the value and use of quantitative data in which I had invested. In my reflections at this time I engaged with sociocultural literature to consider how I might explore sources of invalidity in practice turning my attention to the assessment *process* rather than its products. I established that the use of quantitative data and analyses was not in conflict with a social view of assessment – though it depends very much on the interpretation of the outcomes. As Onwuegbuzie (2002) notes, selecting analysis methods, interpreting statistical outcomes and deciding what outcomes are of practical significance within a research study are all examples of where quantitative data is used in subjective ways. Indeed Erickan and Roth (2006) argue that full

investigations of phenomena need to consider both quantitative and qualitative aspects as *'natural and cultural phenomena are simultaneously quantitative and qualitative'* (2006, p. 16). They propose an integrated approach to educational research in which the questions asked determine the mode of inquiry to answer them. This had a significant effect on my study in that it altered my interpretative stance so that I used the technical analyses to illuminate the issue of 'gradeness' and its stability. Therefore, I used the quantitative analysis to inform my understanding of the meaning of 'gradeness' and to open up the complexity of the problem of conceptualizing examination comparability. I also assumed that it would direct my attention to issues that would benefit from further in-depth qualitative exploration. So although the Welsh dataset offered important technical opportunities, my theoretical reorientation meant that the data and its analysis would serve as a vehicle to inform the more general problem of what examination comparability might or might not mean, that transcended the particular time frame of the data set used. My view after exploring the literature was that using technical treatments as illuminative procedures is compatible with a sociocultural position. The thesis therefore tracks my own development.

I had already found that teachers' perceptions of 'gradeness' challenged assumptions of comparability across subjects and examining groups. I was therefore interested to explore how these perceptions might play out in teachers' assessment practices. There was no anecdotal evidence about stability across sex groups within a cohort for a particular subject and examining group and I intended to consider through my technical exploration whether there were issues here to explore with teachers too. Differentiated assessment schemes for GCSE existed (and still do in 2008). For example different examination papers for the same subject offered access to different GCSE grade ranges (this scheme is called tiering and an examination paper offering access to a specific GCSE grade range is called a tier). This meant that teachers were required to categorize children for examination syllabus and assessment requirements. At the beginning (1995) of the research work investigating the nature of teachers' judgments regarding the choice of syllabus and/or tier of GCSE science papers and consequent influence of these judgments on students' attained GCSE grades was not in the public domain. Even now (2008), there is little work of this nature in the public domain. Professor Burgess and his team at The University of Exeter studied teachers' syllabus and tiering decisions in GCSE mathematics in 2000; Professor Elwood of Queens' University, Belfast continues

to study gendered effects in respect of GCSE tier entry (Elwood, J. 1995, 2001; Elwood, Murphy, Benson, ECER, Dublin, 2005). However, the influential role of teachers in interpreting GCSE syllabuses and in presenting and responding to written coursework was identified by the researchers in the NAS/UWT report of 1989. The literature had already identified potential issues of validity to do with teachers' tiering decisions though this was considered only in terms of entry to GCSE examinations. I was interested to explore further the influences on teachers' judgments and how national assessments at Key Stage 2 (the final four years of primary schooling) and Key Stage 3 (the first three years of secondary schooling), which also functioned through tiered papers, might be implicated in teachers' decision making at GCSE.

My exploration of comparability using a sociocultural approach was informed by my technical analyses and anticipated the mediation of social practices and structures by individuals and considered the interactions between teachers' actions and assessment structures. This is to consider what might lie behind grade distributions that raise further questions about claims of, and meanings attributed to, notions of examination comparability. I was thus interested to explore possible influences within schools in relation to teachers' mediation of the assessment process and hence achievement outcomes (research aim 3). Typically sociocultural approaches which take action as the unit of analysis (Wertsch, 1995) rely on qualitative approaches. I was constrained by the time demands of the dual aspects of my study in relation to the extent to which action could be explored and so decided to do this by probing intentions behind actions. At the commencement of this research in 1995 there were no existing studies of examination comparability using this type of approach. During recent years there has been a growing awareness of the social significance of assessment but policy, public debate and the development of assessment practice still predominately focus around the technical means by which policy is delivered. Therefore, in my study I have an initial emphasis on the technical exploration of assessment outcomes and move to a social exploration of aspects of the process and teachers' practices. This impacts on how I have addressed the literature as first I focus on the technical approaches and studies but clarify the views of mind and of knowledge that underlie them (see Chapters 2 and 3). I elaborate my approach to the technical examination of data in Chapter 3 and my technical analysis and discussion of findings follows this in Chapter 4. In Chapter 5 I present the views of mind and of knowledge that underlie

the qualitative aspect of the study and explain my chosen approach. The qualitative data and analysis is presented in Chapters 6 and 7.

1.5 The wider relevance of this research.

From the very beginning of this research my interest focused on students' performance i.e. attainment in GCSE science subjects rather than how assessment impacts on the learning of science. My interest grew in line with the then emerging debate regarding the comparison of achievements within secondary schools. For example, the use of GCSE examination grades as common currency of achievements was fundamental to the introduction of school league tables, an initiative that evoked considerable criticism from its inception (*TES*, 10 September 1993). School league tables were abandoned in Wales in 2003 but continue to be retained in England as shown in The Daily Telegraph, 27 August 2004, pages 6-7 for the 2004 GCSE results. Certainly by the time of my deliberations (1994-1995) the need to make comparisons of human achievements had become internationally important as evidenced by the rise in examinations that have common currency across continental boundaries. The International Baccalaureate Diploma, which has increased its number of worldwide subscribing schools every year since its introduction in the mid 1960s (IBO: 1994, 2004), is just one such example. *Issues in Setting Standards: Establishing Comparabilities* (Boyle and Christie, 1996), which was published after the commencement of this research, exemplifies the continued growth in international interest in comparability within an examination context.

At the beginning of this research, debates about the interplay of political, social and economic contexts in which national assessment grows and is practiced were informed by the work of Broadfoot (1984) Firestone (1989) and Gipps (1990) amongst others. The notion of national assessment developing in response to society's changing needs from the interactions of different groups of players was (and still is in 2008) generally accepted. Teachers are one such group of players. Teachers are users of GCSE examinations and their acceptance of examination grade awarders' value judgments of their students is significant for the continuing existence of the associated examination system. Students, their parents and selectors in education and employment may also be regarded as users of the GCSE examinations. They are also potentially capable of either accepting or rejecting a particular examination system. For example, the failure of CSE

gaining parity of esteem with 'O' level may be viewed to be in part a result of teachers' and employers' rejection of a CSE grade 1 being comparable to an 'O' level pass for students. By continuing to choose or to reject particular syllabuses and examinations from specific GCSE examining groups, teachers confer or withhold their acceptance of the examining group's GCSE practice. GCSE examining groups are businesses and such candidature movements may stimulate an examining group to review its policy and practice with consequent changes. As a result, I argue, teachers may influence the development of the GCSE examination system and thus national assessment at 16+. The more subtle interaction between assessment, school structures and teachers' judgments and practices in relation to students' GCSE tier allocation has been relatively under-researched. No such work within the context of WJEC GCSE science examinations lies in the public domain. The research of Gorard *et al.* (1999) focuses on the differential achievement of boys and girls in schools in Wales and adopts a technical rather than a sociocultural approach.

The current research includes an attempt to begin to redress that situation. The relevance of such interactions in other GCSE subject domains and other large-scale national assessment systems means that this research has a wider significance for understanding the social dynamic process by which assessment systems develop.

CHAPTER 2

The technical and social dimensions of developing assessment systems: emergent tensions

In this chapter I discuss the psychometric legacy that led to the domination of written examinations and the emergence of comparability as an assessment issue. I consider the theoretical perspective that led to this technical response and how this perspective has been challenged but nevertheless note how the legacy continues to hold sway over practice. I go on to consider the implications of emerging practices for this research.

The thesis explores educational performance first and the assessment process second. In exploring gradeness in external examinations, the thesis focuses on the General Certificate of Secondary Education (GCSE) used to grade students at the end of compulsory education in England and Wales. Contested beliefs about the purposes of education, what constitutes achievement and valid and dependable assessments of this, have served as stimuli for the initiation of changes in the GCSE examination. Tensions arising from these contested beliefs have led to additional changes and have stimulated further transformation of the GCSE examination. My view of the nature of the development of the GCSE examination as a national assessment system is, therefore, one of a social dynamic. An exploration of some of the social and technical dimensions of this dynamic process by which the current form of the GCSE and its attendant tensions have developed clarifies the issues that this research addresses.

An investigation of the history of examinations in the UK is an appropriate first step for understanding the pyschometric roots and the dynamic process by which the GCSE has developed.

2.1 The emergence of written examinations in the UK: high status assessment

Examination systems in the UK developed from the formalization of assessment in the early part of the nineteenth century (Butterfield, 1990). Before that time, for most people with access to education, schooling was important in providing for future lifestyles according to social strata but was irrelevant to the process of occupational selection (Broadfoot, 1996). Social, occupational and personal roles were bound up together and determined largely by birth (ibid.). Assessment, if it existed at all, was essentially a formality.

At the beginning of the nineteenth century there was widespread concern to find solutions to increasing lawlessness in the new industrial cities and to a deficit of workers able to respond to an expanding industrial economy. Many professions also began to feel the need to define and affirm specific levels of professional competence. In an increasingly competitive society these professionals valued the status that would be conferred on them through the creation of a monopoly over a particular profession, the entry to which was controlled according to rigid standards of professional competence. In 1815 the first professional qualifying examinations were introduced by the Society of Apothecaries to ensure that doctors were adequately trained. The implementation of written examinations for solicitors and accountants followed. This growing concern with the affirmation of competence was also reflected in the provision of schooling (Broadfoot, 1996). The Taunton Commission of 1868 called for a system of inspections by centrally appointed school inspectors to see that standards were maintained in elementary schools. The Commission recommended the creation of a central council to administer a non-competitive examination system that would provide a fair assessment of average work (Montgomery, 1965). Although the first inspectors (Her Majesty's Inspectorate - HMI) were appointed in 1840, the recommendation for a central council to control school examinations was not fulfilled until 1917, by which time universities had established control of national examination practice (Broadfoot, 1996).

Rapid growth in local government and expansion of the British Empire in the nineteenth century led to an increased requirement for the recruitment of officers and administrators. Earlier systems of recruitment relying on informal recommendations could not efficiently meet this requirement and became increasingly subject to the criticism that patronage fostered inequity (Sutherland, 1984). In the 1850s Civil Service entrance examinations were introduced, followed in the 1870s by recruitment examinations for military colleges. These examinations were significant as qualifying examinations and particularly for their emphasis on selection (Broadfoot, 1996). They marked the translation of a belief in raising standards by competition from a production to an educational context (Lawson and Silver, 1973). The increasing use of examinations to allocate educational and vocational opportunities represents the origins of both 'criterion-referenced assessment' (the measurement of competencies) and 'norm-referenced assessment' (the comparison of the performances of students to produce a rank order (Broadfoot, 1996)). In a

relatively short time span examinations became predominantly used for selection rather than determining competence.

By the mid-nineteenth century, written examinations had become associated with university education, prestigious professions and positions in the Civil Service. Written examinations were consequently linked with high status (Gipps, 1989) and for this reason were viewed as particularly attractive by the grammar schools. As the industrial capitalist economy flourished in the late eighteenth and early nineteenth century, there was an increasing need for trained workers in managerial positions and the professions. Government needed to encourage a wider range of people to take on these roles to meet the economic imperative (Gipps, 1990). The expanding middle classes realized that education was a means of acquiring social status. These effects led to an increase in the demand for grammar school placements. This expansion was accompanied by anxiety about the quality of schooling. As a consequence the Oxford Local Examinations Board and the Cambridge Local Examinations Syndicate were created in 1857. Both offered written examinations for which either individuals or whole classes in schools could be entered. The age of school examinations controlled by university boards had begun and continues to this day in terms of how standards emerge through the grade demands placed on students to achieve a place at a high ranking University. By the late nineteenth century, the University of London was also administering school written examinations. These examinations became linked to school leaving certificates, thus formalizing an acknowledgement that a particular standard of education had been attained rather than mere attendance at a course. Such certificates became increasingly important for entry to the next level of educational provision.

At this point subjects and achievement were defined by those who 'inspected' (like the Taunton Commission) or controlled the examinations. However, views of the nature of human ability and how to measure it were soon to become influential in these respects.

2.2 Theorising examinations: the emergence of the assessment technicist

Attendance at elementary school became compulsory for children of the working class in England and Wales in 1880. This legislation brought into schools for the first time children who appeared not to be able to cope with its demands, handicapped as they were by physical and mental disability (Sutherland, 1996). The prevailing government system provided schools with grants that were

dependent upon the performance of each child each year in examinations in the '3Rs' conducted by an inspector from Her Majesty's Inspectorate (HMI). The examinations of numeracy and literacy established the norms for achievement for the population based on assumptions that the majority of the population was 'normal' and of similar 'ability'. This system was known as 'payment by results'.

The increasing numbers of students entering compulsory schooling after 1880 who were unable to cope with its demands led to the 'payment by results' system being attacked by several groups, not least teachers and their unions. Compelling children to attend schools that did not meet their needs became an increasingly important matter for public debate. Special schools with additional funding were established to cater for these needs. Society's requirement to identify these children, and provide constructively for them, fuelled a preoccupation with the diagnostic assessment of human abilities (Vernon, 1950) to measure and express how far an individual's abilities deviated from the 'norm' (Sutherland, 1996). Early attempts to do this were very crude, and emphasized external physical features, the 'stigmata' of handicap, for example, *'curved fingers', 'arched palates' and 'lobeless ears'* in Dr James Kerr's investigations conducted in the late 1890s (Sutherland, 1984, p. 21). The limitations of the use of such 'stigmata' for the purpose of classifying children for different types of schools was also a matter for public concern, as illustrated by Dr Francis Warner's comments at this time:

> *It is difficult to define what physical conditions seen, apart from mental tests, indicate the child as unfitted for the usual methods of education.*
>
> (Parliamentary Papers, 1898 xxvi, Defective and Epileptic Children, qq. 752, Warner)

The period from 1870 -1914 witnessed a sustained growth in the technology of testing to address the societal need to match provision to capabilities (Sutherland, 1984, p. 112). Francis Galton was a key figure in this growth. As the first eugenist, his work on mental measurement was influenced by his linking of the two propositions (1) that a struggle for survival takes place in human society as in the plant and animal kingdoms (2) patterns of human reproduction can and ought to be managed. His assumptions were that success, not only in examinations but generally in the world at large, was a criterion of ability and that there was a systemic relationship between physical characteristics, sensory perception and the higher mental processes. He took

measurements of weight, sitting and standing height, arm span, breathing capacity, strength of pull and of squeeze, force of blow, reaction time, keenness of sight and hearing, colour discrimination and judgement of length. These measurements were seen to be of limited value in assessing the mentally defective for selection to schools/institutions since they largely tested simple individual skills and reactions (Sutherland, 1984, p. 53). Nevertheless, Galton's treatment of his findings made significant contributions to the development of the technical aspects of the mathematics associated with mental measurement, contributions that have subsequently served to influence the technical aspects of examinations to the present day. In particular, Galton was the first to apply the Gaussian or 'normal' curve of distribution to the distribution of human abilities. Although the assumption that human abilities were unequally distributed was implicit in all schemes of examinations at the time, few people had concerned themselves with the pattern of frequency of this distribution (Sutherland, 1984, p. 115). Galton argued that there was no reason why the distribution of mental characteristics in a population should not, like the distribution of physical characteristics of that population, follow a Gaussian distribution:

> *There must be a fairly constant average mental capacity in the inhabitants of the British Isles, and that deviations from the average – upwards towards genius and downwards towards stupidity – must follow the law that governs deviations from all true averages.*
>
> (Galton, 1869, p. 32)

The relationships between the distributions of groups of variables were also of central interest to Galton and he introduced the notion of correlation and the calculation of correlation coefficients. Subsequently refined by Pearson, correlation coefficient calculations provided a means of determining if there was a significant relationship between one set of functions and abilities and another in mental testing.

Following Galton, Spearman (1927) went on to develop the uses of correlation in evaluating individual mental tests by means of a matrix of correlation coefficients called factor analysis and as a result, theorized about the nature of mental abilities. Spearman (ibid.) argued for human abilities consisting of two factors; general ability, which he called general intelligence or 'g'; and the residual 's' factor, specific and particular skills. Of these two factors 'g' was considered to be the more important and of principal concern to the mental test constructor. Hence

the less 's' and more 'g' revealed by a test, the better general predictor it was considered to be (Sutherland, 1984, p. 120). Cyril Burt, a leading exponent of the concept of intelligence (Burt, 1921; Eysenck, 1973, pp. 1-22), endorsed Spearman's work (Burt, 1921). Burt used Spearman's approach in his correlation of the abilities of children and their parents to confirm his belief that intelligence was largely inherited.

Sutherland (1984, p. 121) argues that as a result of Spearman's and Burt's work, mental measurement, or psychometrics, based on the theory of natural ability, or intelligence, as an innate and precisely quantifiable general cognitive quality for each individual, was born. There were two major consequences of this new technical assessment paradigm. First, if general ability exists and is predictive of performance, narrow tests of subject achievement are valid because of the assumption that performance across different selections of items would be stable. The invariant notion of ability also reduced concerns about how achievement was assessed. Efficient response modes in terms of time for scoring and reducing marker error were valued as in this paradigm reliability was the dominant issue. Second was the access that test and examination constructors had to sophisticated statistical techniques to consider item behaviour and to establish instruments that behaved according to Guassian assumptions. Shifts away from Guassian distribution were statistically rectified. Test instruments in that sense then constituted the normal curve as opposed to measuring it. The application of the Gaussian or 'normal' curve to the distribution of intelligence in a population, the calculation of correlation coefficients, and factor analysis, all provided a framework, a technical paradigm, for interpreting the results of tests and examinations administered to large populations. Because of the assumptions about ability and performance these interpretations included comparing the results of different tests, and comparing and relating performances of different members of a group. From this point on then the discourse of tests and examinations was characterized by a greater emphasis on technical matters, and the role of those who construct tests and analyse their outcomes, assessment technicists, became increasingly important and distinct, although the Universities were still, and remain in 2008, a powerful force in controlling the content and quality of assessments.

The beginning of the twentieth century saw the emergence of this paradigm characterized by an increasing concern with both the ideological and technical dimensions of testing *en masse* the

mental abilities of children aged 11+. In 1905 the Frenchmen Alfred Binet and Victor Simon were pursuing an opposite hypothesis to that of Galton in respect of what constituted an appropriate test of mental ability in that they viewed intelligence as a complex array of abilities. Nevertheless these abilities were still seen as generic and performance was assumed to function as a surrogate of ability that could predict performance in other domains. They also had a view of normal development in relation to these abilities. They devised a battery of tests that drew on higher mental processes. These tests were judged to measure spatial, numerical and verbal abilities and were graduated by age. The tests were applied to hundreds of Parisian schoolchildren, and Binet and Simon related the results to what the teachers told them about 'normal' and 'abnormal' performances. Subsequently, the tests were revised and then standardized to provide an age-related scale for scoring the results. Binet and Simon's work marked an increase in sophistication of approach to the measurement of children's abilities and their deviation from the 'norm'.

Binet's development of the concept of 'mental age' could be set beside a child's chronological age to provide a way of expressing how far the child's abilities deviated from the norm (Wolf, 1972). Thus a child was said to have a mental age of six years if his performance matched that of an average six-year-old (Sutherland, 1984, p. 54). This simple statement could be enriched by the provision of the individual spatial, numerical and verbal test results. Thus to a limited extent it was believed that a child's individual strengths and weaknesses could be diagnosed. William Stern made it easier for lay people to understand the relationship between mental age and chronological age by his introduction of the intelligence quotient or IQ, seen by some as equal to 'g', general intelligence. Stern divided mental age by chronological age, multiplied the result by 100, rounded off that result to give a single number that could be used to describe the abilities of a child and compare one child with another. Hence the notion of comparability of performance was further strengthened in assessment discourse and practice.

Binet and Simon designed their tests to be applied in a one-to-one situation. As a consequence, in England, the tests were considered to be too costly and labour intensive for widespread use in the selection of children for special schools. However, in England, Winch and Burt had similar concerns to those of Binet and Simon for expressing how far a child's abilities deviated from the 'norm' and they began to experiment with tests on the Binet-Simon model.

Winch and Burt investigated group tests, that is a test that could be administered in written form to a number of people simultaneously and the results scored by an unskilled individual or even a machine. These pencil and paper tests of ability, which could be administered *en masse*, had their answers marked as either right or wrong and in these respects were considered to be objective and marker proof. These tests were based on similar assumptions about the nature of ability but were less extensive than those of Binet and Simon and in that sense did not provide such a rich profile of a child's abilities and because of their mode of administration, did not provide for a dialogue between child and tester and thus serve to influence the child's future learning.

By the 1920s Local Education Authorities (LEAs) in England faced mounting pressure to select children from elementary (primary) schools for placement in secondary schools in as fair a manner as possible so that children from affluent backgrounds were not advantaged. Selection was necessary because a hierarchy of schools existed; grammar schools were generally regarded as better than other forms of state provision but many were fee paying. Examinations for free places in secondary schools were added to the examinations already existing for the award of scholarships and bursaries provided from charitable trusts; often the same examination was used for both. These examinations varied in their nature across the country. Sutherland (1996) reports on how the Northumberland LEA asked Godfrey Thompson, Professor of Education at Newcastle University, to devise tests that could be used as a suitable basis for their secondary school selection procedures. Other LEAs copied these tests and Thompson put the resulting fees into a trust fund. This fund was used to finance the development and publication of further tests (after 1928 called Moray House Tests). These tests were not only group 'mental' or 'intelligence tests' as they rapidly became known, but also consisted of standardized tests of English and arithmetic (ibid.). The use of testing of intelligence for the purposes of allocating children to different types of schools with different curricular provision was further strengthened in the recommendations of the government's educational consultative committees which were published in the Hadow Report in 1926 and the Spens Report in 1938. Sutherland reports (1984; 1996) that in the period 1919 to 1939 between half and three-quarters of the LEAs in England and Wales with responsibility for secondary education used at some point something they called an intelligence test. However, Sutherland (ibid.) comments that LEAs' practice in this respect from 1919 to 1939 often rested on using the

term intelligence test to describe assessments that differed little from their previous qualifying examinations, these being tests of attainment in subjects such as English and arithmetic. Nevertheless, a belief in IQ became widely accepted and intelligence testing became the preferred form of selection for secondary schooling.

By 1938 the 11-plus, consisting of a standardized group intelligence test and examination papers in English and arithmetic, for assessing children's ability for selection purposes was largely in place. When secondary education became free after the passing of the 1944 Education Act, the pressure on the selection process increased. This was because access to grammar school was available for all (Gipps, 1990) and because places were limited as secondary modern schools and technical schools were established for the 'less –academic' children. The 1944 Act did not prescribe the method of selection but made it essential that selection occurred. LEAs had this responsibility and they discharged it by largely relying on the 11-plus.

From the late 1920s onwards, not only were examining techniques becoming more sophisticated (Sutherland, 1994), but the growing research debate about human ability and its measurement, and IQ test construction in particular, served to raise the profile of, and focus attention on, technical matters in assessment discourse (Thorndike *et al.*, 1927, 1933; Vernon, 1979). Scepticism about the claims made for both the concept of intelligence and IQ tests themselves (Vernon, 1979) began to emerge. For example, research had begun to demonstrate that measured intelligence was influenced by environmental factors (Burks, 1928; Hirsch, 1928; Thorndike, 1933) rather than being innate. Alongside this evidence other theories of ability were beginning to emerge (Thurstone, 1938). However, within the UK, intelligence testing continued to be the first response on a national scale to the need to select children for different educational opportunities within finite resources. Such intelligence testing, and the invention of the group test, linked the methods and techniques of the psychology of individual differences into the general discourse on examinations. Thus tests were constructed to perform normatively to reflect the psychological view at the time that human ability was innate, capable of being precisely quantified and normally distributed in the population. These events established the psychometric paradigm for the testing and examination of children and its influence is evident to this day within the UK and world wide. Social practices and structures shape thinking and it is not surprising that decades

of practices which have presumed that 'abilities' pre exist rather than evolve through experience continue to mediate practice today. In addition, educational systems continue to rely on assessment to 'measure' their effectiveness. This means that confidence in outcomes i.e. reliability and the techniques associated with its determination, will continue to be prioritized. Consequently many of the tools and associated practices of the psychometric paradigm continue to be perceived by some, for example GCSE examining group personnel, as the legitimate approach to take when assessment systems are developed. Such people view the notions of quantification, objectivity and 'normal' distribution associated with this assessment tradition as necessary features of legitimate forms of assessment.

Over time society has needed to use assessment for an increasing number of purposes other than selection. As Noah and Eckstein (ibid.) note, reasons for countries legitimizing examinations include their need to:

- select individuals for various stages and types of education, training and employment within material and human constraints;

- secure equity by identifying and rewarding talent so that birth and background are not the sole means to accessing education and employment;

- implement certification of individuals for ensuring that such certificates reflect degrees of learning, rather than simply attendance;

- legitimize knowledge, gain acceptance of a new syllabus and thus influence curriculum reform;

- hold educators accountable for their actions in the sense that examinations may be used to monitor and control educational standards at individual educator, institutional and systemic levels[1].

Arguments about what constitutes an appropriate assessment process have arisen because of differences in views of the purposes that assessment is seen to address and of the nature of the learning-assessment interface, in particular whether ability precedes and predicts achievement or that the two are inseparable and all that we can 'measure' is aspects of the latter (Sternberg 1998). A major conflict of view is evident in the development of national written examinations to accredit

[1] This list is not claimed to be exhaustive. As argued in Assessment: Social Practice and Social Product (Filer, 2000), assessment plays a role in the social structuring of society.

the outcomes of schooling and to provide a basis for selection for higher education during the first half of the twentieth century.

2.3 Expansion of certificated national written examinations in the UK

The early part of the twentieth century witnessed increasing numbers of children completing secondary schooling and seeking higher levels of education. In 1917 and 1922 respectively, the Board of Education introduced the School Certificate which first became a standard school-leaving and second, a university entrance qualification (Broadfoot, 1979). Subjects were grouped in the School Certificate and a pass in five or more academic subjects led to certification. Subjects with written outcomes came to acquire status over other areas of the curriculum, and timed written examinations certified achievement in those subjects (Gipps, 1989). Consequently, timed written examinations became the model for high stakes assessment associated with university entrance certification and completion of secondary schooling.

However, the arrangements associated with the School Certificate generated tensions. Some groups considered the School Certificate to be inflexible and unresponsive to students' individual achievements. For example, if a student reached a satisfactory standard in four subjects they would still not receive a Certificate because five subjects were the minimum qualifying number for certification. As a consequence, the School Certificate was replaced in 1951 with a new end of secondary schooling assessment system, the General Certificate of Education (GCE). In the GCE each subject was individually certificated, and students completing their secondary schooling could do so with varying profiles of GCE subject certificates.

In the GCE system the timed written examination remained the dominant mode of assessment. Pass standards in the different subjects, which were predominantly academic in nature, reflected those of the School Certificate. This led to educational criticisms that the GCE only catered for the top 20 per cent of the 16 + age group school population, leaving many students without any formal recognition of their achievements at school. Whilst the educational arguments foregrounded equity and access, there were also political concerns that the measures of the system were still inadequate to determine its effectiveness. These criticisms and resulting tensions led to the Beloe Report of 1960, which recommended the introduction of other public examinations to provide certification for a further 20 per cent of the 16 + age group. Furthermore, the Report

argued, another 20 per cent of the school population could also gain certification if students were allowed to take fewer subjects than was customary in the GCE.

Even at the time of the Beloe Report (1960), assessment literature continued to focus upon the technical issues of various assessment systems and methods, without paying much attention to the underlying educational assumptions. The treatment of assessment results continued to reflect psychometric concepts (Wood, 1986) with achievement in any GCE subject at this time commonly being reduced to a mark rounded to the nearest five marks as an estimate of an individual's achievements - and potential[2]. This was in spite of increasing criticism of key psychometric concepts associated with ability. For example by the late 1950s intelligence of a fixed and largely inherited nature was disputed as a result of the research conducted by Halsey and Gardner (1953), Simon (1953), and Yates and Pidgeon (1957) amongst others. Their research showed that among other things, social class and environment influenced measured intelligence, and that the measured intelligence of students in grammar schools improved while that of students placed in secondary modern schools deteriorated, in contradiction to the assumption that 'ability' was fixed. Rather than questioning the possibility of measuring intelligence objectively, however, policy concerns focused on providing environments for all students that would be conducive to the development of their intelligence (Plowden Report, 1967). This led to pressure against selection of students for secondary education, which, it was argued, could no longer be justified on the assumption of fixed and differing levels of intelligence (Broadfoot, 1996). This was further fuelled by concerns about the reliability of intelligence testing practice. For example, an NFER study in 1957 showed that 122 students out of every thousand had been wrongly assessed in the 11-plus (Vernon, 1957).

People opposed to the 11-plus and those who were protagonists of theories of learning which challenged Galton's concept of intelligence exerted increasing pressure throughout the 1960s for a common non-selective secondary school system. Such a system, it was argued, would provide for students with different abilities and interests, reflecting a belief that 'ability' was not a fixed, unitary trait. Eventually this pressure resulted in a political commitment to comprehensive

[2] GCE boards differed in how they reported students' achievements, for example the University of London Entrance and School Examination Council reported on a five point scale and Southern Universities Joint Board reported in percentage marks (Wilmott, 1977). After the mid 1970s reporting students' achievements was rationalised to a 1-9 point scale.

schooling (Government Circular 10/65, 1965). Alongside the introduction of comprehensive education, pressure resulting from the recommendations of the Beloe Report (1960) led to an expansion of the national examination system to provide the phased introduction of the Certificate of Secondary Education (CSE) in 1965 to cater for the school population for whom GCE was considered to be inappropriately, academically demanding.

This dual system of examining generated new tensions and sources of conflict, as well as changes in the perception of forms of learning and valid judgements of these. The introduction of the CSE was ostensibly to extend assessment opportunities for the 20 per cent of students deemed to lack the abilities assessed by GCE. It was therefore premised on Guassian terms of normal distribution of abilities. It followed from this that it was possible to place CSE grades and GCE grades on an hierarchical scale. The intention was that some overlap would be ensured so that a ceiling was not placed on students who had the potential to achieve a pass at GCE. The overlap was between CSE grade 1 and GCE pass i.e. grade 6. Because the CSE was targeted at those students outside of what was considered the academic stream i.e. those with the potential to enter higher education, it was established in quite different ways to GCE and in at least two respects. First, the CSE allowed teachers a much greater role in external certification than the GCE. Second, unlike the mainly university-run GCE boards, the CSE boards were regionally based and designed to be teacher-dominated.

Regional autonomy resulted in enormous divergence between the different CSE boards' examination procedures (Broadfoot, 1996). Such divergence was considerably increased by the provision in the Beloe Committee's recommendation of three different modes of CSE examination:

- Mode I, an external examination based on a syllabus devised by a regional board;
- Mode II, an external examination based on syllabuses devised by individual schools or groups of schools and approved by their regional board;
- Mode III, an examination set and marked internally by individual schools or groups of schools on syllabuses devised by their teachers and approved by their regional board.

Modes II and III had the educational benefits of being responsive to local needs. For example my husband devised a Mode 11 syllabus which incorporated a study of aluminium because many of his students' fathers worked in the local aluminium processing factory and educational visits could be

made to aid the teaching of the syllabus content. Many, but not all of the Mode III syllabuses were targeted at specific groups of students, for some of whom the curriculum and assessment requirements laid down for the majority in Mode I schemes were inappropriate (Tattersall, 1994). I taught in a school which introduced a Mode III mathematics for students who were challenged by the computational demands of Mode I mathematics. The Mode III mathematics course was innovative at the time in that it used situations that the targeted students would encounter in their everyday lives and used these to promote the development of simple numerical skills. Thus to a certain extent, in CSE differentiation for certification was facilitated by the creation of different modes of examinations within which, particularly in the case of Mode III, alternative definitions of subject achievement were embedded.

The CSE boards varied in their support for Mode III examining which came to be viewed as a less rigorous assessment system than Modes I and II (Broadfoot, 1994). Consequently, the CSE boards varied in their provision for enabling teachers to meet the curricular and assessment needs of all of their students for whom the CSE examining system was designed. Furthermore, despite the intention for the CSE to certificate different kinds of achievement and encompass subjects that had previously been accorded relatively low status in the curriculum, the reality was different. Nuttall (1984) comments that CSE examinations like their GCE counterparts were dominated by timed written examinations with a preponderance of questions requiring essay-writing skills and factual recall based on syllabuses appearing to serve a university entrance model rather than the needs of those entering work at 16. Although work carried out by students throughout a period of time (coursework), for example the production of a piece of woodwork, was assessed in the CSE, such assessments were regarded as lacking in status (Butterfield, 1990). As long as the CSE had to compete for credibility with the GCE, it was bound by many of its traditions and methods (ibid.).

2.3.1 The emergence of comparability as a technical concern

The widespread reorganization of schools along comprehensive lines increasingly exposed the weaknesses of the dual system of examinations (ibid.). Administratively, schools a found it demanding to meet the requirements of at least two examining groups. Consequently, the CSE examining system had hardly begun before attempts were being made to devise a common system

of examining at 16+ (ibid.). A concern about comparability of grading standards was another reason that a common system of examining at 16+ was advocated (Wilmott, 1977). The term grading standards is taken to refer to the award of a particular grade by an examining group for a given level of performance on the part of an examined student (ibid.). In 1966 fourteen examination groups were administering the CSE examination and there was some doubt as to whether there was equivalence of grading standards between them. Whether or not there was parity of grading standards between GCE and CSE also became a matter for debate. Although the top CSE grade was linked to an 'O' level GCE pass in order to accredit the former, employers and parents never accorded the CSE the status that had been hoped for it (Wilmott, 1977; Gipps *et al.*, 1986).

The debate over comparability of grading standards led to the Schools Council commissioning the Examinations and Tests Research Unit (ETRU) of the National Foundation for Educational Research (NFER) to investigate standards in the CSE examinations. The aims of the resulting four studies (Schools Council, 1966; Skurnik and Hall, 1969; Skurnik and Connaughton, 1970; Nuttall, 1971) were (i) to see the degree to which the CSE boards were able to agree with one another as to the recognized standard in a number of subjects, and (ii) to attempt to relate the CSE grade one/two boundary to the GCE 'O' level grade six / seven (pass / fail) boundary (Willmott, 1977).

The findings in general indicated that some variation existed between the CSE boards in the recognition of subject grading standards in any one year. However, this variation was not excessive and in only one case was a consistent deviation noted for the same examining group and subject over the three examination years (1966-68) in the study. Difficulties were experienced in investigating the level of the CSE grade one/grade two boundary in relation to that of the GCE 'O' level six/seven boundary (Nuttall, 1971). The available evidence indicated that it was more difficult to obtain a CSE grade one than a GCE 'O' level pass in a number of subjects (ibid.), which ran counter to beliefs that the CSE was less rigorous and lacked credibility compared with GCE.

The initial years of the co-existence of GCE and CSE examinations therefore brought to the fore the issue central to this research, namely the problem of comparability of standards in examinations. Just as the role of teachers in national assessment was radically changed by the

CSE, so too was that of the assessment technicist, largely examination board / group personnel and researchers. From this point in time assessment literature revealed that examining board / group personnel and researchers of the technical aspects of examinations paid increasing attention to the issue of comparability of grading standards in 16+ (and 18+) national examinations.

Throughout the 1960s and 1970s teachers and other groups exerted increasing pressure for a common examination system at 16+ so that all students could follow the same syllabus for entry to the same examination.

> *At present, schools have to make difficult decisions on the selection of pupils* [students] *for GCE or CSE courses, perhaps as early as the end of the third year ..... Early choices cannot allow for the development of pupils' abilities ..... many teachers would say that the task ..... is one which they find particularly difficult and unrewarding ..... this practice may, to some extent, pre-empt the results of the examinations themselves.*
>
> (Schools Council, 1975, p. 9)

2.3.2 Differentiation within a common examining system

The Schools Council presented its proposals for a new 16+ common examination in 1971 (Schools Council, 1971), and in 1974 trial examinations were taken by nearly 70,000 students, with 1980 as a target date for introduction (Broadfoot, 1996). In these trials examining board / group personnel identified difficulties in examining students with a wide range of ability with a single examination (Schools Council, 1971). Such a single / common examination contained questions for the more able which could not be attempted by the less able; questions appropriate for the less able did not draw on the higher order skills of the more able students. A common examination for all students was seen as inefficient in terms of the time required for students to respond and marking time. It was also seen as demotivating for students to be faced with questions that were overly challenging. Furthermore, for 16+ examinations, the co-existence of CSE and GCE laid the foundations for a tradition of students with different 'ability' being taught in different teaching groups using different syllabuses associated with the different examinations. It was recognized that a common 16+ examination was feasible to administer within time constraints but was unlikely to be able to discriminate adequately across the ability range concerned or that school structures and resources

could cater for such an assessment approach. Consequently it was considered that a *system* of examining rather than a single examination might be more appropriate (Tattersall, 1983).

The word 'differentiation' did not appear in the Schools Council Report of 1975 but this was at the heart of the various suggestions. The major concern of the then incoming government was that a lowering of standards could result from a common examination incorporating differentiation (Gipps *et al.*, 1986) as it might be associated with the CSE which had not achieved parity of esteem with GCE. However, with standards tied to GCE for students' examination at 16+[3], there was increasing governmental support for a system of different examination papers catering for different groups based on their assumed 'ability'. Differentiation thus became a common feature of the proposals for a new 16+ examining system. This commitment brought fresh emphasis to the issues of comparability of grading standards and differential examination difficulty in assessment discourse and debate.

Government concern for these issues amongst others led to the formation of yet another committee, the Waddell Committee, to conduct another feasibility study of a common examining system at 16+. The Waddell Report of 1978 (DES, 1978) echoed the 1975 Schools Council Report recommendation for a single system of examining. This recommendation came to fruition in 1986 with the introduction of the General Certificate of Education (GCSE), and its first examination in 1988.

2.4 A theoretical shift

The introduction of the GCSE followed an ongoing debate on equity in relation to the construct and consequential validity of assessment. The former drew on theorizing that challenged the view of mind assumed in the psychometric paradigm. The growth of the psychometric paradigm coincided with attempts to discover general laws about learning that would lead to a scientific theory of learning and teaching. These theories were premised on a belief in an objective reality. Teaching was to engineer the appropriate stimuli whereby knowledge from outside was brought into the mind. Fundamental to this view of learning and of knowledge was that there was no process of meaning making on the part of the learner, and knowledge was assumed to be stable across people (Bredo, 1999). Mind was viewed as an information processor, passively receiving information and

[3] GCE Ordinary Level (GCE 'O' Level) was the examination taken by students at 16+.

acting on it. These learning theories were, however, subject to increasing criticism based on experimental evidence that suggested that intrinsic motivation mattered in learning. Piagetian theorizing, hitherto neglected, which placed action and self-directed problem-solving as central to learning came to the fore in the 1960s as it helped explain the significance of learning for its 'own sake' (Wood 1987). Piagetian theorizing also, through the articulation of age-related stages of development, helped explain why humans learned particular things at certain times and moved attention away from biological determinism towards notions of 'readiness to learn'. However, whilst Piagetian theory, as taken up in education, replaced the image of the passive mind with that of the constructing mind it provided no challenge to normalizing practices as the Piagetian stages were regarded as: *'natural normalised stages of development towards scientific rationality'* (Walkerdine, 1989, p. 198).

Constructivist theorizing developed in the 1970s and 1980s, that built on the central Piagetian notion that meanings were not given but constructed by mind, moved away from notions of normalized staged development and began to challenge the basis of both the psychometric paradigm and norm-referenced assessment. To meet this challenge the new examination required a reconsideration of how performance would be referenced. Further challenges to assessment procedures raised by this shift in understanding of human achievement involved the assessment instruments and methods. Within the psychometric paradigm a narrowly defined assessment instrument was appropriate given that this was understood to be predictive of general achievement. However, a constructivist view of mind assumes that individual achievement has no ceiling and can vary within and across subjects. Hence a profile of subject achievement is the goal of constructivist assessment. To establish a profile there is a need to specify more carefully the domain being assessed to identify the key concepts, procedures and skills associated with it and for instruments to be representative of these specified domains.

A constructivist model of mind however, whilst rejecting an objective external reality, does not necessarily pay attention to the social mediation of learning (Cobb, 1999). What it is concerned with is 'fitness for purpose' and hence assessment instruments had to more carefully reflect the contexts of use of the functional knowledge assessed. This opened up opportunities for a wider array of assessment methods to be used and reduced the emphasis on examinations which were

seen to be suited to the assessment of only certain aspects of subject knowledge. The assessment reforms described represented in part a shift away from psychometric paradigm towards an educational assessment culture (Gipps, 1994) which presumes that many achievements are attainable by all students, but how and when they will attain them will vary.

2.4.1 The mediation of assessment policy: the political agenda

The educational agenda for change in the form of national examinations coincided with a political agenda to enhance the accountability of the system and as part of that to move towards increased centralization of the curriculum and its assessment. Since the late 1970s and into the 1980s the government funded a national monitoring programme, the Assessment of Performance Unit. This monitored the performance of representative populations of 11, 13 and 15 year old students in core subjects in England, Wales and Northern Ireland. However, the information available was not at student or school level. In a political climate where education was implicated in the economic and social well-being of a country, national monitoring was seen to be an inadequate system for accountability purposes. The Waddell Report (DES, 1978) recommended that as the new 16+ common examination system, the GCSE should provide national criteria for subject titles and syllabuses. The change in referencing to national criteria was seen politically to meet the challenge of comparability essential for a system that provided an accountability framework by which schools' performance could be measured. The GCSE national criteria were intended to ensure a degree of comparability among the syllabuses produced by different examining groups (Murphy, 1986) and *'to enable the grades awarded to be accepted with confidence by those concerned'* (DES, 1978). The creation and subsequent development of the GCSE as the first 16+ common national examination system in the UK brought the issue of comparability in examinations to the very forefront of assessment discourse and debate. Furthermore, comparability was to be further enhanced by a significant reduction in examining groups with the proposal to have the system managed by three or four regional consortia combining both GCE and CSE boards. This signaled the move towards more centralized control of the system.

The creation of national criteria also represented an unprecedented detailed specification by the government of the objectives of learning, syllabus content and the types of assessment device to be used for all courses of study in several of the major subjects of the curriculum (Nuttall,

1990). Historically, prior to this, the universities shaped the academic school curriculum by their domination of GCE examinations. On the other hand, groups of teachers usually decided the syllabus content of CSE examinations. As Broadfoot (1996) has argued, the power and influence exercised through the examination system was, historically speaking, dispersed between many different groups.

In the same year as the first examination of the GCSE with its national criteria the Education Reform Act (ERA) of 1988 legislated for the establishment of a curriculum (National Curriculum) for all students of compulsory age in maintained schools in England and Wales with the knowledge, skills and understanding that they were deemed to need for adult life and employment (DES/WO, 1989). The ERA (1988) also established a national assessment system further strengthening the government's control of the assessment process within primary and secondary educational sectors. It gave unprecedented control of public examinations[4] to the government through statutory advisory bodies. First, the National Curriculum Council (NCC) and the School Examinations and Assessment Council (SEAC) were established for advising on respectively curricular and assessment matters. Since that time these two bodies have been merged into one, the Schools Curriculum and Assessment Authority (SCAA), which by 1997 changed again to become the Qualifications and Curriculum Authority (QCA). In 2004 the National Assessment Agency (NAA) was launched as a subsidiary of QCA with responsibility for the delivery of national tests and examinations. All of these bodies have served at various times to advise on the appropriateness of syllabuses for 16+ national examinations. Without their recommendation, a syllabus cannot gain qualification status. GCSE Regulations and Criteria (SCAA, 1993a), which guide the approval of syllabuses, became statutory in their application for the 1998 examinations. In accordance with Section 5 of the 1988 ERA, the Mandatory Code of Practice for the GCSE (SCAA, 1993b) also explicitly reinforced the requirement for quality and consistency in the examining process across all examining bodies. The code of practice was seen as the means to ensure that grading standards were constant in each subject across different examining consortia and different syllabuses and from year to year.

[4] The Schools Council had advocated an increase in *teachers'* governance of the new 16+ common examination. Broadfoot (1996) comments that the Council proved largely impotent to influence policy in practice. After its demise in 1982, successive Secretaries of State increasingly exercised their power.

Thus, the ERA (1988) provided measures to enhance comparability of examinations.

One could argue that the redistribution of power contingent upon the introduction of the GCSE national criteria and the ERA (1988) giving the locus of control of the school curriculum and national assessment to the government would enhance comparability in 16+ national examinations. Controls included the requirement of:

- detailed frameworks setting out a common core content for each GCSE subject's syllabus;
- detailed assessment objectives and methods for each GCSE subject;
- specified weighting in the associated marking schemes of the GCSE subjects;
- the moderation by examination groups of compulsory assessment by teachers of coursework in the GCSE subjects.

These requirements might reasonably be expected to increase uniformity of GCSE examination practice within and between examining groups and enhance consistency in grading standards between different examinations.

2.4.2 Referencing systems

Criterion-referencing with grade-related criteria was originally an objective for the GCSE (Joseph, 1984a, 1984b) but remains an illusive goal (and in 2008). Alongside an educational argument, concerns for comparability of examination grades between different syllabuses and between different examining groups created support for criterion-referencing with grade-related criteria being applied to the GCSE (Orr and Nuttall, 1983). The interplay between the proponents and antagonists of criterion-referencing linked to grade criteria for the GCSE exemplifies the technical and social dimensions of developing national assessment systems. Of relevance here is a discussion of the referencing systems that awarders use to decide the relative merits of, and thus the grades awarded to, students' examined work. The referencing systems' terms are often confused (AEB, 1995). Conventionally, criterion-referenced tests are meant to measure the degree of competence attained by a particular student on a profile of attainment (Glaser, 1963). In such tests the assessment domain is specified in detail and interpretations of the student's performance are made against a profile of possible attainment for the assessment domain. As stated in Glaser's seminal paper on criterion-referenced testing:

> *Measures which assess student achievement in terms of criterion standards thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others.*
>
> (Glaser, 1963, p. 520)

Criterion-referencing assumes a conventional numerical scoring process at the level of individual questions. However, the questions are selected so as to be representative of the assessed domain. In that way the score obtained by a student can be interpreted as that student's expected attainment on the entire domain (hence the need to use a well-defined domain in criterion-referencing). A major intention of conventional criterion-referencing is to provide formative information. In summative forms it is intended to provide users of assessment information with an understanding of what students know and can do.

In its original sense, norm-referencing means standardizing, i.e. identifying each student's test or examination score within the distribution of attainment of the student's peers as in the intelligence tests discussed earlier in this Chapter. Conventionally, and in contrast to criterion-referenced tests, norm-referenced tests do not specify the assessment domain in detail. The questions in norm-referenced assessments are not representative of the assessment domain as a whole, although they are assumed to be predictive measures across domains. The rank ordering provided by norm-referenced testing, *'only indicates an individual's success in relation to their peers and not in terms of the knowledge, skills and understanding achieved by that individual'* (Murphy *et al.*, 1996, p. 62).

The terms criterion- and norm-referencing are often assigned meanings other than their conventional ones. With the introduction of the National Curriculum in England and Wales the alternative meaning of criterion as *standard* came into use. Brief verbal statements acting as standards in terms of particular competencies were developed and called statements of attainment, and applied across the 5-16 curriculum at Key Stages defined by age (discussed later in the chapter) (DES, 1989). This approach is often termed 'strong criterion-referencing' because of the strength of the descriptive inferences about students' attainments that it claims to make possible (Cresswell and Houston, 1991). When this approach is used, verbal descriptions replace numerical scores. Concerns about the precision of meaning of such verbal descriptions and the accuracy of their

application have been voiced (Sadler, 1987; Ruddock *et al.*, 1993; Wolf, 1993), particularly in the context of the assessment of the National Curriculum in England and Wales. Furthermore, controlling the representativeness of the questions so that a student's performance on them may be interpreted as the student's expected attainment on the whole assessment domain does not occur in strong criterion-referencing in contrast to its conventional counterpart.

As Christie and Forrest (1981) have argued, national examinations usually have a clearly discernible assessment domain and, in the GCSE question setting process, effort is made to ensure that the examination as a whole represents it effectively. Assessment grids giving a breakdown of the rationale used for sampling the syllabus in the construction of the papers are often provided by examining groups in their syllabus regulations (an assessment grid is shown in Appendix 1). GCSE syllabuses do not define assessment domains with the precision required by conventional criterion-referencing. The question papers are related to broad areas of knowledge, so that there is potential for considerable variability in relation to the constructs being assessed year on year and between examination consortia. In this sense the construction of the examinations is based on strong criterion referencing. Examining groups provide brief descriptions of performance by grade, and what distinguishes them, with their syllabuses. However, in awarding grades the GCSE places an emphasis on numerical scoring of questions, uses aggregation of component numerical scores and rank-orders students' total numerical scores for a particular syllabus. An approximate normal distribution of total numerical scores is expected and proportions of students falling into numerical score ranges are also anticipated to approximate those of previous years. In terms of an emphasis on quantification and a normal distribution of testing outcomes the norm-referenced traditions of the psychometric paradigm are evident in these aspects of the GCSE.

Grades are not, however, arbitrarily assigned to numerical mark ranges in the GCSE. Grade awarders bring value judgments to the grade awarding process. These value judgments are supported by reasons (Fogelin, 1967; Beardsley, 1981) based upon the use of tacit standards held by them as a 'guild of professionals' (Sadler, 1985, 1987, 1989). In this evaluative process, marks are considered in terms of how they represent the guild's (ibid.) view of the value of particular grades in that assessment/subject domain. Of relevance here is the notion that the GCSE cannot simply be regarded as norm-referenced because it has aspects that emphasize rank-ordering of

numerical scores with anticipated approximate normal distributions. For this reason, national examinations such as the GCSE are often erroneously said to be norm-referenced.

The nature of the assessment referencing system for GCSE might have been different if the Schools Council and the government of the 1980s had had their way. Prompted by a concern that grades in national 16+ examinations were not comparable across different syllabuses, the Schools Council in 1979 strongly recommended the development of national agreed definitions of standards of work in these examinations. By 1980 the government was instructing the examining groups to begin work on defining national 16+ examination grades in terms of performance:

> *Consideration should also be given to the possibility of incorporating (in the national criteria) some elements of criterion-referencing of grades, or some grades in the 7-point scale. This might help certificates to be more informative for users about the things candidates* [students] *have shown they can do and go some way to free the award of grades from statistical norms of quality or performance change over time.*
>
> (Department of Education and Science, DES, 1980)

In the conventional sense a grade criterion is an attribute to be assessed (Christie and Forrest (1981). The Schools Council (1979) and the DES paper of 1980 regarded it as a *standard* based on a view that grade criteria should be written statements that prescribe the level of attainment required to justify the award of a particular grade (Murphy, 1986).

Murphy (ibid.), commenting on the developments subsequent to the DES (1980) document shown above, argued that the challenge of grade criteria is thereafter avoided. First it was avoided by the Joint Council for National Criteria, a body set up by the GCE and CSE examining groups / boards. This Council redefined the task that they had been given by the Department by distinguishing between 'criterion-related grading' and 'grade descriptions' (Joint Council for National Criteria, 1981). Under the former, students would be required to demonstrate predetermined levels of competence in specified aspects of the subject in order to be awarded a particular grade. This was the intention in the Scottish 16+ common examination / certification system examined for the first time in 1986 and is referred to later in this section. Grade descriptions, on the other hand, are a different matter: they merely attempt to indicate the levels of

attainment likely to be shown by students awarded particular grades in a subject (ibid.). Gipps (1990) argues that this redefinition was prompted by the Joint Council's concern about the technical problems associated with a national grade related examination system. Whether or not this was true, the DES also came to accept the more limited aim of producing grade descriptions:

> *Grade descriptions, as outlined above, may prove to be a step towards a longer term goal. The Secretaries of State have asked the boards* [GCSE examining groups] *to set themselves the objective of making the award of all grades conditional on evidence of attainment in specific aspects of a subject.*
>
> (DES, 1982)

As Murphy (1986) notes, nothing more was heard of grade criteria for a couple of years until, in a bold last-ditch attempt, the Education Secretary, Sir Keith Joseph, tried again to inject them into the final stage of the development of the new examination because he saw this as improving standards in schools. He viewed the use of grade criteria in criterion-referenced assessment as being supportive of positive achievement, a central tenet of the government's proposals for the new examination. Examination grades equated to absolute standards of competence, skill and understanding for the attainment of students of different abilities would facilitate teachers and students working towards new targets (Joseph, 1984a, 1984b). Scotland was ahead of England and Wales in this respect. Changes in the 16+ examining system took place in Scotland for a limited number of subjects with the first criterion-referenced examinations taking place in 1986. The intention was to provide more useful information about students' achievements for both students and their teachers. However, in attempting to identify the content and skills that students could be assessed on and the different degrees of mastery that might be demonstrated, the complexity of the resulting system made it unworkable, prompting a radical simplification. Popham (1987) warns of this scenario.

Given the Scottish 'experience' of grade criteria, it is hardly surprising that Sir Keith Joseph's enthusiasm for its speedy introduction for the GCSE was not shared by those charged with its development, the Secondary Examinations Council (SEC).

> *The rigorous specification of full criterion-referencing for assessment in the .....* [new common examination] *would result in very tightly defined syllabuses and*

patterns of assessment which would not allow the flexibility of approach that characterizes education in this country.

Nevertheless Council agreed with the DES that a move towards a greater degree of explicitness was desirable ...

(Secondary Examinations Council, 1984, p. 2)

In the run up to the examination's implementation, DES publications became more vague about when grade criteria would emerge:

... the proposed examination will be designed, not for any particular proportion of the ability range, but for all candidates [all students taking the examination], *whatever their ability relative to other candidates, who are able to reach the standards required for the award of particular grades. Grade criteria are being developed for this purpose and will be incorporated into the subject criteria and*

syllabuses as soon as practicable.

(DES, 1985)

Secondary Examination Council (SEC) circulated draft grade criteria for consultation prior to the implementation of the examination but the time frame for comment and subsequent redrafting was very short.

Thus, as the GCSE first became a reality for teaching purposes in 1986 and subsequently for examination in 1988, grade criteria were still in development. Syllabuses were linked to grade descriptions for individual subjects. The grade descriptions describe a representative attainment worthy of the grade, rather than the attainment of every student awarded the grade and cannot, for the reasons discussed above, be used as criteria for judging the attainment of all students. They merely 'convey the flavour' of a grade (Wilmut and Rose, 1989).

Criterion-referencing with grade criteria prompted by various groups' aspirations to:

- provide more valid and useful information about students' attainments;
- facilitate the setting of learning targets;
- provide a means of ensuring greater comparability of grading standards between different syllabuses and assessment domains/subjects by various examining groups,

remains an illusive goal for the GCSE. These aspirations emanating largely from the educational agenda concerned with construct validity and the social justice of assessment raised challenges for the system's technical dimension that could not be met. Consequently the technical dimension dominated emerging practice and ensured that the issue of central concern to the current research, comparability of grading standards, was, and continues to be, problematic and a significant concern in assessment discourse and debate.

2.5 Practices within the national system of assessment: issues for the research

2.5.1 The psychometric legacy

GCSE became the 'measure' of the final output of compulsory schooling in the education system. However, if the system was to act on assessment information there was a need to have measures at points within the system where action could be taken at school and teacher level both within the school and by government. The national assessment system set up by the ERA (1988) was also to provide parents and carers with the information they needed to evaluate the effectiveness of the provision received by their children. The National Curriculum was taught to students in specific age ranges which were referred to as Key Stages, each Key Stage (KS) and age range being respectively KS1(5-7years), KS2 (7-11years), KS3 (11-14years) and KS4 (14-16 years). The curriculum content for each subject was delineated into Attainment Targets. Each of these consisted of Statements of Attainment which in turn were categorized into a hierarchy of difficulty referred to as Levels 1-10. Different ranges of levels were assigned to each key stage as follows: KS1 (levels 1-3), KS2 (levels 2-5), KS3 (levels 3-7, 8 in mathematics) and KS4 (levels 4-10).

The national assessment system provided for students to be assessed at the end of each Key Stage. Initially a system of standard assessment tasks (SATs) combined with teacher assessment of students' work over a period of time was envisioned. The SATs were to be performance- type assessments allowing for a range of response modes and style in keeping with an educational view of achievement which gives priority to validity. These tasks proved difficult to develop and administer on a national scale to every student and their reliability was brought into question. In addition, teacher assessment was controversial and subject to criticisms about its reliability in such a high stakes form of assessment. Consequently, national curriculum tests, referred to as standard tasks at age 7 and standard tests at age 11 and 14, were introduced. These were based on strong

criterion-referencing and descriptions of standards in the statements of attainment in line with GCSE practices and their development and marking was, and continues to be in 2008, administered by a variety of funded bodies including examining groups. The assessment procedures and the instrument development is overseen by the National Assessment Agency. The standard tasks for KS1 students at age 7 are no longer administered largely due to parental and carers' concern about the pressure they placed on young children; assessment at this Key Stage was removed as a requirement. At KS2 and 3 parents and carers receive the standard test results and the teacher assessment in English, mathematics and science. These national test results are available in the public domain by school and it is the test results that are used by Government and the media to comment on 'standards' of performance year on year and against which schools are held to account (similar arrangements currently exist in 2008 except for Wales, which has withdrawn from the national test arrangements at Key Stages 1 to 3 and is piloting a new Key Stage 3 assessment system).

At KS4, continuous assessment through GCSE coursework gave teachers some role in the final assessment, although this was (and continues to be in 2008) moderated externally by the GCSE examining groups to monitor standards. Outcomes of the GCSE examinations are reported as grades (A-G until 1994: A*-G from 1994), and across subjects and examining groups are used as a common currency of achievement. GCSE examination results of schools in England and Wales in relation to the number of students achieving five A*-C are routinely published and used to provide rank orders of schools (league tables) as indicators of the effectiveness of their educational provision (this practice ceased in Wales in 2001). The appropriateness of such practices for revealing the effectiveness of secondary schooling is questionable (Nuttall *et al.*, 1989). For example there could be a learning zone (special needs unit) attached to a school and the students' results in the zone / unit are taken into account in the calculation of the school's five GCSE A*-C grade results. This was the practice during the time Wales participated in league tables of schools based on their GCSE results. Certainly the associated statistical problems are subject to debate (Guskey and Kifer, 1989; Goldstein, 1991, 1995). Outcomes at other key stages are also published in league tables by the proportion of the population achieving the level expected of students, for example level 5 at KS3. Although not a concern of this study in relation to comparability, similar

assumptions about the common currency of levels obtain for key stage assessment as it does for GCSE. However, in England league tables continue (2008) to be used as a means of holding schools and LEAs accountable for their students' educational outcomes and providing a means by which a fall or rise in standards might be measured. GCSE performance i.e. the proportion of students achieving 5 grades A* - C as the final output measure are seen as particularly significant and it is the measure used by government to determine whether a school is failing, which can lead to closure. It is also one of the indicators that influence parental choice.

Thus whether or not it is valid to do so, GCSE examination results are now used extensively in a *comparative* way for the purposes of monitoring educational effectiveness and accountability. These practices and the debates concerning their appropriateness have strengthened the importance of the issue of comparability within a national examination context.

2.5.2 Differentiation practices

Tiering

Another key feature of the national assessment system that is significant in discussions of comparability is the approach to differentiation adopted across the system. Again, how this aspect of the system developed was influenced both by an educational and a political agenda, the former being concerned with students' experiences and the latter with enhancing the efficiency of a very costly accountability framework that reported at the individual student level. The approach adopted in the GCSE influenced practice at the key stages.

GCSE was introduced with the claim that it would enable all students *'to show what they know, understand and can do'* (DES, 1985). In making this claim the Secondary Examinations Council stressed that assessment should be a positive experience for all rather than a dispiriting one for some, and therefore students should not be presented with tasks that were too difficult (SEC, 1985). Allowing students to show what they could do rather than presenting many students with tasks which they were likely to fail became known as 'positive achievement' and was facilitated by differentiation – pitching papers and questions at different levels of difficulty. The use of differentiated papers in GCSE reflects this view of achievement.

Commenting on this feature of the GCSE Nuttall writes that differentiation:

> *... has come to have a very specialized meaning within the context of assessment: that is, that the* [GCSE] *examination system, though distinguishing between seven different grades of performance, should at the same time differentiate between students in a manner that allows every student to demonstrate in positive terms what they know, understand and can do.*
>
> (Nuttall, 1990, p. 144)

Tattersall (1983) wrote extensively about possible models of differentiated examinations. These she described as:

- all students take a common examination paper together with one of a number of additional papers of varying levels of difficulty;

- all students take a basic examination which tests all facets of a course except the content and skills which are deemed appropriate for only the most able students;

- different students take overlapping papers testing overlapping syllabus levels of difficulty each of which is designed for a subset of the ability range.

Tattersall (ibid.) commented that different subjects with their associated pedagogical practices may suit different models of differentiated examinations. The GCSE National Criteria (DES, 1985) required some subjects to be examined using differentiated papers. This was (and continues to be in 2008) the case for science, the subject central to this research. The model most frequently adopted in GCSE subjects including science is where different students take overlapping papers testing overlapping syllabus levels of difficulty. This form of differentiated examination papers is often referred to in GCSE as tiering.

> *Tiering provides pupils* [students] *with the opportunity to show what they know, understand and can do by presenting them with question papers that are targeted at a band of attainment.*
>
> (SCAA, 1996, p. 3)

In tiering students entered for a GCSE in a given subject sit different examination papers according to their teachers' expectation of their likely performance. In tiered examination papers the grades available to students are limited by a 'ceiling' and a 'floor'. For example in a model of a two tiered system such as:

Higher tier	Grades A* to D are available
Foundation tier	Grades C to G are available

Grade D is the 'floor' for the higher tier and grade C is the 'ceiling' for the foundation tier. The easier assessment route may lead to the possibility of reaching only a grade C. Grade C is judged to be the pass level at GCSE. Conversely, the harder assessment route may allow students to reach grade A*, but may permit only a grade D as the lowest level; if the student fails to get that, then she / he usually gets nothing rather than a grade E as a consolation prize (exceptional grades are discussed in Chapter 3)

Tiering was and continues to be used in the national standard tests administered at the end of Key Stage 3 (11-14 age group) of the National Curriculum. In science there were two tiers of tests, a lower tier covering levels 3-6 and a higher tier for levels 5-7, and an extension paper to give students entered for the higher tier access to a level 8 award, though this is no longer in use in 2008. Tiering introduces a number of issues about comparability. First is the assumption across the system that strong criterion-referencing allows levels and grades to be used as common currency. Second, that different papers with overlapping questions can be used to determine grade or level performance on the assumption that the examinations and tests are representative of the domain. These assumptions are, as the literature indicates, questionable because of the limitations of the criterion-referencing approach used and other mediating factors. Since its introduction the model of differentiated examinations used in GCSE subjects has varied with time. Indeed, between 1998 and 1999, just before my engagement with teachers in the qualitative part of my research, tiering arrangements between different subjects and examination groups were standardized. The majority of GCSE subjects, including science, adopted a two tier model this pertains in 2008.

These GCSE and Key Stage 3 tiering arrangements had implications for my research. First, I would need to take decisions about which tiers of GCSE science examination papers I would include in my quantitative investigation of examination performance (Chapter 3). Second, in my qualitative investigation the mediation of prior assessments on teachers' decisions in relation to their students' tier allocation for GCSE science examinations was considered significant.

Certification routes

The period, 1989-1995, of constant curricular and assessment change was also marked by a proliferation in the type and number of GCSE examination syllabuses and this is particularly true for science. With the rationale of providing *different routes* to certification for students of varying abilities, the GCSE examination groups produced Single Award Science, Double Award Science and Triple Award Science GCSE syllabuses. All had to conform to the GCSE Regulations and Criteria for Science (SCAA, 1993; SCAA, 1995) which provided instructions for syllabus aims, assessment objectives, syllabus content, schemes of assessment and assessment techniques, and grade descriptions for grades F, C and A. Some of these syllabuses took a linear form, others a modular; some were labeled co-ordinated, some integrated; some emphasized a particular pedagogical approach (see Nuffield and Salters). This plethora of syllabus types at this time further enhanced the issue of comparability of grading standards in assessment discourse and debate.

In the early 1990s it was possible (pertains in 2008) for schools to 'shop around' for their GCSE science syllabuses. It was not uncommon for a school to enter their students for GCSE Triple Award Chemistry with one examination group and GCSE Triple Award Physics with another (WJEC, 1994). Schools would also enter some students for GCSE Triple Award science syllabuses with one particular examining group and other students for GCSE Double Award Science with another (ibid.). This phenomenon was prompted by a belief that it was easier to obtain a high grade in some syllabuses than in others (OUDLE, 1994; SEG, 1994; WJEC, 1994). From 1995, the last of my three years' of quantitative examination performance data collection, students presenting themselves for the GCSE Triple Award in biology, chemistry, physics were required to take all three of three science subjects with the same examining group. Nevertheless, this phenomenon of schools 'picking and mixing' GCSE science syllabuses from those made available by the different examining groups still occurred and triggered a rise in interest about comparability of grading standards by the School Curriculum and Assessment Authority. Inter-group comparability exercises (SEG, 1995) were instigated in response to concerns about this issue.

As users of the GCSE examination able to be selective about their choice of examining group and syllabuses, teachers have acquired a significant influence on GCSE development.

Determining the degree and nature of that influence has been relatively under-researched. This further convinced me of the usefulness of exploring the nature of the qualitative judgements teachers make when entering their students for GCSE science examinations.

2.5.3 Continuous assessment

The move heralded by GCSE to view achievement in broader terms and to recognize that individual achievement has no ceiling and can vary within and across subjects also led to a concern with how achievements were assessed. This reflected a concern with the validity and not the reliability of assessment. If mind is constructive and has to create meaning and negotiate it then the circumstances in which assessments are made matter. So assessing practical skills by paper can inform to a degree, but knowing that students know what type of measurements are needed to meet a particular purpose is more validly assessed in the context of use. This approach was labeled as the 'fitness for purpose' of assessment. In the pilot of Key Stage 1 assessment of science for example, the intention was to use a practical context throughout. This, however, proved administratively difficult and was soon abandoned. Subsequently all key stage assessment has been by written form and response only.

In responding to 'fitness for purpose' at GCSE the intention was to introduce an element of continuous assessment called coursework. Coursework was defined by the Schools Curriculum and Assessment Agency (SCAA) in 1995 as consisting of:

> *in-course tasks set and undertaken according to conditions prescribed by an awarding body. Coursework activities are integral to, rather than incidental to, the course of study. Coursework is normally marked by a candidate's* [student's] *own teacher according to criteria provided and exemplified by the answering body, taking national requirements into account. It is moderated by the awarding body.*
>
> (SCAA, 1995a, p. 13)

Coursework was introduced into the GCSE to allow a fuller representation of achievement including objectives that are difficult to assess in timed written examinations, such as practical and oral work. It allows teachers to assess their students' achievements. For this research coursework is significant because like tiering and syllabuses it has undergone significant change in terms of its

nature and significance in GCSE assessment. It therefore has significance in relation to comparability over time

The amount of GCSE coursework has varied across subjects and within syllabuses for the same subject right from the GCSE's first examination in 1988. For example, my son was able to take GCSE English with 100 per cent coursework in 1990; coursework in mathematics was not introduced until 1991 when syllabuses with 20 per cent were the most popular. The year 1991 also saw John Major, the Conservative Prime Minister, speaking out in favour of traditional testing and against teachers' assessments in GCSE and the Key Stage 3 national assessments. The then political agenda focused on raising standards and the proliferation of different GCSE syllabuses with different coursework component percentage weightings was seen to undermine this (Baker and O'Neil, 1994). This led to the devaluing of coursework as the continuous assessment element in national assessment and its reduction in GCSE in 1994. Coursework became limited to 20 per cent in most syllabuses, with mathematics retreating to include some syllabuses with no coursework and those for English being limited to 40 per cent, half of which was allocated to oral assessment. At a time when research was increasingly showing the social nature of learning, so the GCSE (and other national assessments) retreated from this model and view of mind that were at least in part its original *raison d'être*.

After John Major's call for a retreat to examinations away from teachers' assessment of coursework, research began to provide evidence that girls' average coursework marks were higher than those of boys and more 'bunched'. Girls were shown to do better on coursework relative to examinations (Stobart *et al.*, 1992). In a more detailed study, Elwood (1995) compared the intended weighting of assessment components with the achieved weighting and showed that coursework functions differently for boys and girls, examinations playing a more important role in final grades than intended for girls than for boys.

So during the time of the examinations relevant to my quantitative data (1993 – 1995 inclusive) and my engagement with teachers in my qualitative investigation (1999 – 2000), there were significant changes in GCSE coursework arrangements. At the same time research was beginning to reveal evidence of boys and girls' differential performance on coursework and written examination components. The nature of the coursework arrangements and any changes therein

across the examinations relevant to my quantitative study are therefore significant to aid the interpretation of my findings.

2.6 Concluding remarks

This Chapter has argued that the GCSE is neither a norm-referenced nor a criterion-referenced assessment system but one in which for social, political and technical reasons there is both a statistical treatment of numerical marking and an evaluation of students' attainments based on human value judgements. The emphasis on the statistical treatment of numerical marking in the GCSE reflects a psychometric approach to assessment. Although the GCSE has raised the visibility of assessing different kinds of human achievement, like its predecessors it continues to reinforce aspects of an assessment culture that is concerned with selection, curriculum control and hierarchies of kinds of knowledge in which theoretical achievement is accorded a higher status. Resulting tensions generate a variety of concerns.

A concern for *comparability* has consistently emerged from this Chapter's discussion of developing 16+ national assessment within the UK. Over time this concern has shifted in focus, for example from comparability between different assessment systems as in the case of GCE and CSE, to comparability between different syllabuses administered by different examining groups. My view is that during the 1990s comparability of grading standards between different GCSE subjects emerged as a central focus of concern (and remains so in 2008). Arguably, this is largely due to GCSE grades from different assessment domains/subjects being increasingly used as common currency in the monitoring of educational effectiveness.

CHAPTER 3

Examination comparability: approaches to its investigation and my quantitative research design

When I first came to this research it was within a quantitative paradigm. This was largely due to the influence of existing examination comparability studies and to a lesser degree, my training as a physical scientist, both of which had ontological and epistemological positions aligned respectively to realism and positivism. These influences are now discussed for how they shaped my first investigation of examination comparability. Some of the challenges to investigating examination comparability within a quantitative paradigm which I encountered are also considered. Finally I discuss the research design for my quantitative investigation of comparability.

3.1 Examination comparability studies: an overview

Two observations emerge from a review of previous comparability studies. First, over time a common approach to such studies has been consistently adopted. This common approach is arguably without a theoretical foundation (Goldstein, 1986; Cresswell, 1997) and is only just beginning to be challenged theoretically. Second, the level of research activity in the field has varied significantly over time with a peak in the 1970s, a fallow period in the 1980s and resurgence of the issue in the late 1990s.

The common approach has been to regard comparability within an examination context as a *technical* problem with a *technical* solution. Certainly this was the approach adopted throughout the 1970s - a period of high research activity in the field largely because of concerns about GCE and CSE examination comparability. This particular approach is underpinned by the assumption that it is possible to establish quantitatively equivalent levels of attainment across qualitatively different assessment domains. This assumption was challenged (see Goldstein, 1986) by theories of educational measurement. However, no acceptable alternative approach was forthcoming and consequently there was no large-scale British research on comparability published from the mid-1970s until the mid-1990s.

During the early1990s social and political concern about the ongoing significant decline in 'A' level science entries stimulated interest in the relative difficulty of 'A' level subjects. Fitz-Gibbon and Vincent's (1994) study of 'A' level grade outcomes in 1994 was the first published

study of different subjects' grade outcomes since the 1970s. This study was concerned with establishing the value added by schools, between GCSE performance at 16 plus and 'A' level performance at 18 plus. Hence it is not specifically relevant to my research on examination comparability but interestingly, this study adopted the common, technical approach to comparability in its statistical treatment of different subjects' 'A' level grade outcomes. It was not without its critics (Goldstein, 1996; Cresswell, 1996), although the criticisms were of the specific technical procedures rather than the technical approach.

Since 1994, in the UK there has been a legal requirement that national examination results are published as indicators of the success of individual schools. This practice has given a new importance to defining comparable standards across subjects because different combinations of subjects taken by different schools can affect the rankings of those schools in the published league tables. It is largely for this reason that the period from mid to late 1990s and early 2000s is associated with a resurgence in comparability research. For example, advances in multilevel modelling techniques by Goldstein (1995) reflect a new emphasis on factors that can potentially influence the outcomes of different examinations. Thus the significance of a technical approach to comparability is that it abounds in practice and informs teachers', parents' and students' beliefs.

3.2 Comparability as a technical issue

The majority of comparability studies focus on statistical aspects of examination grade outcomes. In technical studies comparable within an examination grade context does not mean that an individual student taking different examinations would necessarily attain the same grade in those subjects (Goldstein, 1986). As discussed in Chapter 1, it is generally accepted from an educational view of achievement that individuals develop understanding and skills differentially across subject domains. They may also reasonably be expected to respond differently to different examination papers sat over a period of time, as is the case in GCSE administrative arrangements. In technical studies of national examinations comparability is an issue to do with group rather than individual performance. The Schools Council's Forum on Comparability in the late 1970s noted in relation to comparability:

> *... the expectation is that had a group of examinees followed another board's syllabus and taken its examination, they might reasonably be*

expected to have obtained the same average grade.

(Schools Council, 1979)

Technical studies have extended this view of examination comparability beyond a consideration of the average grade of a group of students to a statistical analysis of students' distribution of grades. Comparable standards of grading are assumed to have been applied if there are similar grade distributions for groups of students taking two different examinations.

A fundamental assumption is implicit in this view of comparability. If examination grade distributions are understood to provide a reliable indication of the comparability of grade standards, then it follows that grade distributions depend *only* upon those standards. However, I argue that this is unlikely to be the case as examination grades reflect the interactions of many different variables and influences.

3.2.1 Variables influencing examination grade outcomes

The examination grades achieved by any one student may be influenced by many factors. Some of the key factors are grouped below as different types of variables, although the list is not definitive:

- features of the syllabuses (*syllabus variables*):
- features of the examinations (*examination variables*);
- aspects of the students' schools (*school variables*);
- characteristics of the students (*student variables*);
- social factors both within and outside of the examination process (*social 'variables'*),

although the appropriateness of labelling such features, aspects, characteristics and factors as discrete variables is questioned – they are more appropriately referred to as influences, as discussed in more detail later. The effect of these 'influences' on examination achievements is enormously complex because of their variety and the ways in which they may interact with one another.

Some of the variables may be regarded either as artefacts of assessment or as factors which interact with these artefacts of assessment. For example, mode of response is an *examination variable* and an assessment artefact. Students' sex or socio-economic class (*student variables)* are known to interact with this artefact (Murphy, 1982) to affect the achievements that students can demonstrate, though the process of how this is achieved is a social one and not within the remit of this thesis for exploration. Sex is a variable; gender emerges in interaction between people and is a

social context influence rather than a 'variable'. Its effect can however be considered as influencing the interactions of students with assessment contexts and tasks and therefore mediates outcomes (see Gipps and Murphy, 1994, and Murphy and Ivinson, 2004, for a discussion). Such assessment artefacts and interacting factors are, in my view, sources of invalidity that challenge assumptions about examination comparability. This view is explored in the rest of this section.

Defining comparable grading standards only in terms of identical grade distributions assumes that the syllabuses upon which the examinations are based define assessment domains that are attributed with the same profile of cognitive demand. It could be that assessment domains *do* require different cognitive demands as a default position (Gardner and Hatch, 1989). For example the domain of physics could be claimed to make more quantitative demands than other domains pertinent to this research, such as biology, which arguably may preclude comparing grading standards between the two subjects.

Another view (Christie and Forrest, 1981) is that examination standards can only be comparable if the syllabuses define assessment domains, which are appropriate to the particular subject at a particular level of education. Less clear is *who* decides what is 'appropriate' – the examiner drafting the syllabus, a government agency with an overview of the control of such syllabuses or the users of these syllabuses such as teachers? Or is the decision the result of the composite of these different groups' effects? This notion of value-laden judgements about the knowledge associated with assessment domains is in opposition to the default position of Gardner and Hatch (1989).

My own view is that there *is* a default position: different assessment domains *are* associated with different cognitive demands. This view is based on my experiences of teaching and examining GCSE science subjects. As a result I judge physics, for example, to be inherently more quantitative than biology. In my view different constructs are assessed in biology, chemistry and physics. Consequently, I do not interpret different grade distributions for these assessment domains as necessarily implying a lack of comparability in grading standards. Rather the assessed students may have interacted differently with the domains' associated cognitive demands.

However, I also argue that there is evidence of a process of 'valuing' in syllabus and examination paper construction. For example the outcomes of human judgements of value (*social influences*) are evidenced in:

(i) how examination syllabuses even for the same assessment domain may emphasize some cognitive demands more than others;

(ii) how examinations may contain a disproportionate representation of some cognitive demands through selection of items by the person(s) constructing the examination papers.

It is possible for example for two physics syllabuses to differ in the extent to which they emphasize the quantitative dimension of the domain. This may in part be due to the approach to criterion-referencing discussed in Chapter 2 which allows significant variation within a criterion description. The production of the same grade distributions from the examination of these two syllabuses could be interpreted as evidence of comparable grading standards, although the domain is significantly different in the attainments measured. As reported in Chapter 1 (WJEC, 1995), some science teachers view examining groups' GCSE science syllabuses and papers as differentially enabling their students to show what they know and can do. Such teachers may choose one examining group's syllabus and associated examination papers in preference to those of another on the basis of their judgements about their learners.

In the early 1990s I entered my students for WJEC GCSE chemistry papers. My counterpart at a neighbouring school entered his students for the Midland Examining Group's (MEG) GCSE chemistry papers. In his view, his students did less well with WJEC because his students' level of literacy skills disadvantaged them on WJEC's papers with their emphasis on continuous prose answers. In contrast I viewed MEG as disadvantaging my students because the associated syllabus and examination papers used contexts that tended not to be in my students' everyday experience. However, we were both identifying construct irrelevant variance, that is variation in performance due to factors other than what is assumed to be assessed in the mark scheme (Messick, 1989). In my counterpart's view the construct being assessed was altered by the WJEC examination papers' language demands, whereas for me it was altered by MEG's choice of question context. As teachers we were both mediating the grade outcomes of our students by our

choice of 'appropriate' examinations. The production of the same grade distribution for our two different examinations could be claimed to be indicative of comparable grading standards – we would disagree.

The subject content, which is the knowledge that is specified in a syllabus (*syllabus influence*), may affect the perceived facility of the course and hence the motivation of the students (*student influence*). There may also be variation in the amount of organizational detail offered in different syllabuses: even within WJEC's GCSE Biology, Chemistry and Physics syllabuses this was the case at the beginning[1] of my research. For example, in the 1994 GCSE examinations pertinent to this research, the physics syllabus presents a detailed breakdown of the relationship between the assessment objectives and content in terms of mark allocations awarded to knowledge/recall, understanding and processes. In addition, there is a breakdown of content areas and their weightings (Appendix 1). The GCSE Biology syllabus for the same year does not offer as much detail in terms of mark allocations and types of skills (ibid.).

Arguably the physics syllabus makes the subject more accessible in the degree to which it offers guidance to teachers. Similarly, differences in syllabus demand are also likely to affect the motivation of the students and potentially their attained grades. Interactions of this type make it impossible to distinguish between the effects of differences between organizational features and the effects of differences in cognitive demand of syllabuses. Comparability studies involving the statistical analysis of grade distributions ignore such differences and assume that the effects of syllabuses upon teaching and learning are identical for a population/cohort of students.

The nature of examination tasks (*examination influences*) may influence students' achieved examination grades. The examination conditions that Nuttall (1987) suggested were conducive to eliciting students' best attainments included:

(i) tasks that are concrete and within the experience of the individual student;

(ii) tasks that are clearly presented;

(iii) tasks that are perceived as relevant to the current concerns of the student;

(iv) conditions that are not unduly threatening, for example sufficient time is allowed for task completion.

---

[1] This is less true for syllabuses currently in use for examination in 2008.

Arguably, for (i) and (iii) what is concrete, within the experience of the student and perceived as relevant for the student's current concerns will vary from student to student (*student influences*) and illustrates the potential for examination and student influence / 'variable' interaction (Gipps and Murphy, 1994).

> *The conclusion is inescapable: ... Assessment (like learning) is highly context-specific and one generalises at one's peril.*
>
> (Nuttall, 1987, p. 115)

Even when grade distributions from different examinations are identical for their associated examination populations, they may not be identical for well-defined sub-groups within these (*student influence*). Differences between girls' and boys' achieved GCSE grades are well established and depend to some extent at least on differences in the assessment techniques used in the examinations. For example Newbold and Scanlon (1981) explored the relationships between boys' and girls' performances in a range of subjects, including biology, offered by the University of Cambridge Local Examinations Syndicate (UCLES) in their 1979 examinations. All of the examinations studied contained multiple–choice components. They concluded that:

> *In line with earlier findings in individual subjects and the sciences, there seems in all five subjects to be a pattern, which associates the relative success of boys with objective and semi-objective test forms, and girls' relative attainments with tests requiring a large degree of free response.*
>
> (Newbold and Scanlon, 1981, p. 5)

Murphy's consideration of the examination statistics for GCE 'O' level for England and Wales for the period 1951 – 1977 also showed that boys were advantaged by multiple-choice type questions (Murphy, 1982). This finding was replicated in his study of the Associated Examining Board (AEB) GCE 'O' level science results for 1976 – 1979. Stobart (Stobart *et al.*, 1992) has shown coursework to advantage girls. Other studies (Quinlan, 1991) have also shown that this is not necessarily the case, for example in subjects where coursework takes the form of continuous assessment within lesson time, as in science: such differential attainment is also shown to be dependent upon the nature of the coursework task (Cresswell, 1990).

Gipps and Murphy (1994) were the first to evaluate international evidence in an attempt to examine the extent of observed group differences in assessment performance and understand what these might reflect. They asked to what extent were apparent differences in achievement created by particular approaches to subject knowledge and the way in which it is tested: can changing the structure and content of the test change the pattern of results? In the case of boys and girls in particular Gipps and Murphy's evaluation of the research indicates 'yes', it can. These analyses provide overall differences between sex groups but understanding gender as an influence that emerges in social interaction challenges these gender effects being consistent across a sub-group. Rather they emerge for some girls and some boys and depend on interacting factors within an assessment situation including the experiences, identities and expectations that students bring to them (Murphy and Ivinson, 2004). An overall difference indicates the presence of a gender effect but not which individual students are affected.

Students' perceptions of the 'difficulty' of a particular subject and its associated examination can vary from student to student, subject to subject and examination to examination. For example students' perceptions of the 'difficulty' of a subject may be affected by the associated examination / assessment arrangements:

> *"English language is easy because it's coursework ... you write your essays and the teacher picks out the best ones for the exam board ... there's no hassle of an exam like in maths."*
>
> (Rhys Clewer, Year 11 student in 1994, personal communication)

Such perceptions may influence students' confidence and motivation. It is widely accepted that affective factors such as these mediate students' performance in assessments (Stobart *et al.*, 1992). These affective factors may interact with numerous sub-group and test variables (Gipps and Murphy, 1994). Sub-groups alone may be variously defined in terms of ethnic origin, socio-economics and gender and all have been shown to be influential in examination performance (Smith and Tomlinson, 1989; Nuttall *et al.*, 1989; Drew and Gray, 1990, 1991; Troyna, 1991). Such student: social: examination interactions illustrate the complexity of making comparisons of examination performances. However, these interactions have largely been ignored in the interpretations of findings of technical studies of grade distributions.

School variables affect students' GCSE grade outcomes. A school's GCSE examination entry policy is recognised as being a school variable that influences students' achievements (Cresswell, 1997, p. 73). Schools wishing to enter students who have a highly developed knowledge of science may prefer to use a particular GCSE examination because of its syllabus (*syllabus variable*). A grade distribution skewed towards high attainment would be a reasonable expectation of such a scenario. A lack of similarity with the grade distribution from another examination could say more about schools' different student entry policies than about grading standards for the respective examinations.

Tiering, a model of differentiated examination papers discussed in Chapter 2, is another assessment artefact that mediates school entry policies. The 'ceiling' and 'floor' effects on available grades in differentiated papers make it vital that teachers enter their students for appropriate tiers. Research shows that choosing the appropriate tier of entry for students is problematic (Good and Cresswell, 1988d; IGRC, 1993; Gillborn and Youdell, 1998). Tier entry decisions are based on teachers' knowledge of their students. The range of grades available to a student depends on both the student's performance and their teacher's judgement of them for tier entry (Wiliam, 1996). Differential performance between boys and girls was argued by Stobart *et al.* (1992) as being influenced by tier entry schemes. They reviewed teachers' comments from surveys and case study interviews and found that more boys than girls were entered for the foundation tier in a three tier model used in GCSE mathematics. Disaffection with GCSE mathematics was seen by teachers as being greater for the boys than the girls placed in the foundation tier. Girls were seen as being more content than the boys to take a lower tier. The greater disaffection shown by lower attaining boys influenced teachers' decisions about whether to enter them at all for the GCSE examination. In contrast more girls than boys were entered for the intermediate tier with its maximum grade B. Stobart *et al.* (1992) suggest that the bigger female entry in the intermediate tier reflects an underestimation of girls' mathematical abilities by their teachers who perceived girls as being less confident and anxious about failure. Teachers responded by entering proportionally more girls than boys for the intermediate tier which avoided the risk of being unclassified if performance dropped below grade C. Able girls' lack of confidence and boys'

abundance of confidence was seen by teachers as a factor affecting performance (Stobart *et al.*, 1992).

The research by Gillborn and Youdell (1998) also suggests that tiering introduces additional barriers to equality of opportunity for students from different ethnic origins, and in particular Black students. Black students were more likely to be entered for the foundation tier and less likely to be entered for the higher tier, and the most significant inequality of access to high grades was in those subjects which operated with a three tier entry model. Teachers tended to be cautious when entering their students for tiers and 'played safe' to avoid students falling off the floor of the top tier. Such tier entry effects may result in some examinations being skewed in their grade distributions. A foundation tier may have its results skewed away from the lower grades because students capable of higher grades have been inappropriately entered for this tier paper. It might then be assumed that the foundation paper has been inappropriately 'easy'. Simply comparing grade distributions from different examinations ignores inappropriate tier entry effects.

This section has highlighted the seemingly intractable nature of making valid examination comparisons. Consequently, a consideration of the methods used in previous studies is necessary to inform the methodology for my research.

3.2.2 The technical approach: treatment of variables

A review of completed technical examination comparability studies reveals two methodological issues. First, the majority of the studies control for only one type of variable. Generally the variable relates to students and, much less commonly, to schools. Second, the studies differ in the type of student variable that they prioritize for such treatment. In addition, many of the studies are located in the psychometric paradigm and their assumptions are based on the view that ability is fixed, and performance on assessments reflects and predicts this ability. Thus it is assumed that performance across subjects for individuals should be consistent; further, if ability can be measured it can be controlled for when comparing examination grade outcomes.

Controlling for student variability

Studies may restrict comparisons of grade distributions from two different examinations to the same group of students (Nuttall *et al.*, 1974; NEAB, 1993; UCLES, 1993). The assumption is that each student has a fixed level of ability and because of this should achieve the same grade on each of the two examinations if the examinations have similar severity of grading. It is assumed that

affective factors such as motivation and confidence, and the effects of syllabuses and examinations upon students' examination performances are identical and therefore controlled for, simply because the same students are involved. Forrest and Shoesmith (1985) in their review of inter-board examination comparability studies conducted by examination boards during the period 1978-1985 state that there is no reason why this should be so.

> *If we consider the entire group of candidates* [students] *taking both Physics and Chemistry, say, in a particular board, how does the distribution of grades in Physics for that group compare with the corresponding distribution in Chemistry? It is sometimes argued that, if everything else is in order, the two distributions should be the same since the population of candidates is the same; but there is a counter-argument that too little is known (for example, about the degree of motivation and intensity of study in different subjects) to justify such an assertion.*
>
> (Forrest and Shoesmith, 1985, p. 11)

Nevertheless, the control of student variables by investigating the same group of students is used annually by examining groups in a method known as subject pair analysis[2] for their own internal monitoring of the validity of their measures.

I approached the GCSE examining groups in early 1994 requesting information about the methods used for investigating examination comparability. The six groups that responded confirmed they used subject pair analysis and revealed a general reluctance to place the outcomes of such analysis in the public domain. One group, the Northern Examinations and Assessment Group (NEAB), expressed their misgivings about the validity of the outcomes. This group provided an example of subject pair analysis from their 1993 GCSE science examinations (Table 3.1). The students' GCSE grades have first been converted to integers (A=8, B=7 down to U=1). The mean grade gained in each of the two subjects by all of the students having taken that particular pair of examinations is then calculated and the difference obtained. Overall, using this method Table 3.1 is said to show that Chemistry syllabuses A and B are more severely graded than those for either Biology or Physics A and B.

---

[2] See Nuttall *et al.*, 1974 for a detailed account of subject pair analysis.

Table 3.1 NEAB 1993 GCSE Subject Pair Analyses

Mean Grade (A)		Mean Grade (B)		Difference (A – B)*
Biology	6.5	Chemistry A	6.3	0.2
Biology	6.5	Chemistry B	6.1	0.4
Biology	6.7	Physics A	6.7	0.0
Biology	6.8	Physics B	6.4	0.4
Chemistry A	6.6	Physics A	6.7	-0.1
Chemistry A	6.4	Physics B	6.5	-0.1
Chemistry B	6.5	Physics A	6.9	-0.4

*Value is *positive* when examination A is less severely graded than examination B

*Value is *negative* when examination A is more severely graded than examination B

NEAB stressed that the subject pair method outcomes do not represent a definitive statement on the relative severity /leniency of particular syllabuses but merely serve as one of a number of indicators. Factors identified by NEAB that could mediate the validity of subject pair analysis outcomes included:

(i) the possible differences in the teaching of the subjects paired together; the lengths of the courses; the amount of time devoted to the subjects; the disparity in school facilities *(school variables)*;

(ii) the possible differences in the interests and motivations of the students in the paired subjects *(student variables)*;

(iii) students who took only one subject are omitted from the analysis. The proportions of students taking only one subject varied from one subject to another. Consequently the method only partially represents the grading by subject, which affects the validity of the data in Table 3.1.

Concern (i), involves school variables that affect the students' learning opportunities and not in my view, the comparability of the examinations. Concern (ii) reiterates my concern with student variables expressed earlier. Concern (iii) illustrates the complexity of the limitations of adopting the 'same student' treatment of student variables. However, it is argued that the assumptions upon which the 'same student' methodology is based become more tenable as the number of students increases. When a large number of students are involved there is more chance of a similar spread of examination entry policies, and some affective factors such as motivation and other variables upon which the grade distributions in the compared subjects depend. The difficulty lies in trying to establish the population size needed to justify the assumptions. Furthermore, sub-

group effects may skew overall examination performances. For example, science examinations emphasizing electrical content set in contexts not reflective of girls' out-of-school experiences could alter some girls' confidence and actual examination performance (Johnson and Murphy, 1986). Thus, simply due to the girls' sub-group effect, the same group of students could produce different grade distributions in two different science examinations. By taking larger numbers of students the effects from socio-cultural factors such as gender on examination performance would become even more evident.

It is interesting to note that in my 1994 communication with GCSE examination groups, the only other source of subject pair analyses outcomes was from UCLES and these were only reported in different sex sub-groups (Table 3.2). In contrast to the NEAB analyses, UCLES converted their GCSE grades to integers on a scale where A=1, B=2 down to U=8. From Table 3.2 one might say that for both boys and girls the severity of grading increases in the order of Biology, Physics and Chemistry.

Table 3.2 UCLES 1993 GCSE Subject Pair Analyses

BOYS

Mean Grade (A)		**Mean Grade (B)**		**Difference (A-B)**
Chemistry	2.647	Biology	2.385	0.262
Chemistry	2.685	Physics	2.371	0.314

GIRLS

Chemistry	2.933	Biology	2.435	0.498
Chemistry	2.485	Physics	2.381	0.104

*Value is *positive* when examination A is more severely graded than examination B

*Value is *negative* when examination A is less severely graded than examination B

When I communicated with the GCSE examining groups again in 1995/6 they confirmed that subject pair analyses were still used annually in examination comparability studies for internal group use. NEAB alone supplied some of their analyses outcomes, this time for 1995 examinations and with reference to sex groups. The various pairings of the subjects Biology, Chemistry and Physics revealed that there were no apparent differences in severity of grading between these subjects. However, the performances of the two sexes showed that girls were significantly ($P \leq 0.05$) more severely graded on Chemistry than Biology but there was no difference in grading for the boys on these two subjects. When Biology and Physics were paired, girls were significantly ($P \leq 0.05$) more severely graded on Physics and boys were significantly ($P \leq 0.05$) more severely

graded on Biology. Similarly, when Chemistry and Physics were paired, girls were significantly ($P \leq 0.05$) more severely graded on Physics and boys were significantly ($P \leq 0.05$) more severely graded on Chemistry. Thus the similarity in subject mean grades mask underlying differences in the mean grades of sex sub-groups.

These findings further exemplify the limitations of the 'same student' methodology in terms of sub-group effects. In Johnson and Murphy's (1986) APU research, which unlike subject pair analysis was not based on psychometric theory, sub-group effects were anticipated and indeed, were identified. The effects were considered to be due to the result of interaction of the sub-group, for example boys and girls, though not necessarily all boys and all girls, with particular aspects of the examination. The assumption then that differences in performance outcomes for boys and girls across subjects reflect differences in grading is challenged as they might reflect differences in boys' and girls' views of what is significant in the item or indeed differences in their opportunities to learn – in short, reflecting the social gender mediation of teaching and learning. A combination of demands in assessment artefacts can differentially affect students' performances and arguably to an extent that reduces the validity of comparing grade outcomes from different examinations taken by the same group of students.

The 'same student' treatment may claim to control several variables that affect examination performance, but this claim does not hold in reality. At best, and on condition that the compared examinations are taken, ideally by the entire age cohort, but more realistically within the context of GCSE by very large numbers of students, the 'same student' treatment and subject pair analysis is only an indicator of examination comparability.

Controlling for ability

This treatment, like the previous one, is based upon a view of ability as innate general intelligence which is predictive of achievement; an achievement is viewed as being predictive of subsequent achievements. This approach claims to take groups of students with the same distributions of ability, and assumes that all other student, school, examination, syllabus and social variables are identical for these groups. The approach not only assumes that general intelligence predicts subject performance but that when differences in the spread of the examination population's general intelligence are taken into account, any remaining differences between examination grade

distributions in different examinations are indicative of a lack of comparability between the examinations themselves (Bardell, Forrest and Shoesmith, 1978; Forrest and Shoesmith, 1985).

One could argue that controlling for ability is less sophisticated than using the 'same student' approach on the basis that it only identifies two variables as being significant, general ability and achievement. However, the 'same student' approach's claim to control for many variables is, as discussed, questionable. Taking students with the same distributions of innate intelligence rather than taking the same students is only more sophisticated from a psychometric perspective. Nevertheless, this treatment was most frequently used in the 1970s by Schools Council researchers (Willmott, 1980). This method controlled for intelligence using a reference ability test such as the NFER's scholastic Aptitude Test 100. The use of such a test provided information about an examination population's spread of general intelligence rather than about specific skills and types of knowledge (Bardell, Forrest and Shoesmith, 1978).

For example the 1968 CSE Monitoring Experiment (Nuttal, 1971) involved samples of students who sat CSE or GCE 'O' level in summer 1968 being additionally tested with the NFER's scholastic Aptitude Test 100 in the preceding February and March. The GCE rather than the CSE findings are used to illustrate Nuttall's work, as in terms of performance in national 16+ examinations, the GCE examination population is more like the GCSE populations used in the current research i.e. it involved the top 20% of the entire 16+ population. For each student in Nuttall's sample, their total score on the NFER's Aptitude Test 100 and the grades achieved in as many of the ten subjects (art, biology, chemistry, English language, English literature, French, geography, history, mathematics, physics) constituted the raw data. The subject grades at this time were in integers so that the smaller the integer, the better the student's performance. When the average test score for students was plotted against their average grade in each subject it was found that groups of students with the highest test scores tended to be those with the smallest average grades. Some sort of difference between the subjects existed and the extent of this difference was investigated using the same regression method as used by Nuttall (1971) in his 1968 CSE Monitoring Experiment.

A standard measure of the ten subjects was first obtained, defined as the average of the average grades awarded in each GCE subject and the average of all the Aptitude Test scores. This

measure provided the average relationship between the GCE grades and the Test scores across all ten GCE subjects. It could be used to predict the average GCE grade that would be expected for any given Test score if grades in each GCE subject were awarded using the same standards. For example, the average Test score of the chemistry students in the sample for GCE board 2 was 55.0 (3.5 points better than the average Test score of the complete sample). The regression method predicted that the corresponding average chemistry grade for an average Test score of 55.0 should be 5.11, on the assumption that grades in chemistry were awarded on the same standard as grades in the other nine subjects. As the mean grade actually awarded in chemistry was 5.44, chemistry was identified as being severely graded by 0.33 grades (5.44-5.11 = 0.33). This process was repeated for each subject in turn and the results are shown in Table 3.3.

Table 3.3 Sample estimates of mean grade severity in GCE board 2

Regression method (Nuttall *et al.*, 1974)

Subject	*Estimate of severity*
Art	- 0.49
Biology	- 0.14
Chemistry	0.33
English language	- 0.49
English literature	- 0.24
French	0.25
Geography	0.09
History	0.16
Mathematics	0.12
Physics	0.37

Positive values indicate a tendency towards severity of grading, while
Negative values indicate a tendency towards leniency of grading.

Physics and chemistry are interpreted as being more severely graded than biology and this trend was replicated across all of the GCE boards included in the study (ibid.). Nuttall *et al.* (ibid.) identified bias in the reference test (the NFER Aptitude Test 100 used in the study) rather than differences between the subjects as explaining the lack of comparability of grading standards between the subjects. He argued that the nature of the items in the Test 100 was such that those students entered for mathematics or for science subjects would obtain significantly higher scores on the Test than students in other subjects simply by virtue of their having followed mathematically orientated courses (ibid.). In other words Nuttall claimed that the Test scores for the different groups of students might not be directly comparable. In this sense he challenged that 'intelligence'

existed or could be measured arguing that the reference test itself was just another knowledge and skills test.

This highlights a fundamental problem in using such a reference test in comparing examination grading standards. There is a theoretical incompatibility between the reference test that assumes norm-referencing against general intelligence and examinations that seek to measure developed knowledge and skills related to specific subjects and involving strong criterion - referencing. The reference test assumes a psychometric view of achievement whereas GCSE examinations with their strong criterion-referencing in syllabus and examination paper construction reflect an educational assessment perspective with many achievements being attained by all students. Using 'ability' measures to investigate examination comparability implies that examinations are inappropriate for the uses made of them. Nuttall and Willmott imply as much by suggesting in 1972 with their call for a 'single general intelligence test' in place of public examinations.

Controlling for students' attainment relevant to the different examinations being compared

To investigate comparability between different subject examination performances it seems more theoretically compatible with a constructivist perspective on educational achievement using strong criterion-referenced assessments to use a reference test which itself measures the distribution of students' developed subject knowledge and skills i.e. subject attainment than one for 'general ability'. By controlling directly for subject attainment the influence of a large number of variables that would affect it appear to be stabilized. However, this treatment introduces another problematic assumption – that the test by which subject attainment is measured is itself comparable with each of the examinations being compared. The difficulties of constructing such a test that could be independently shown to be equally relevant to the different assessment domains of the compared examinations caused the ending of the Schools Council comparability studies which used subject attainment tests as controlling instruments in the1970s (Forrest and Shoesmith, 1985). Even when the subject attainment reference test is constructed from common elements of the compared examinations, the same problems arise (Newbold and Massey, 1979).

Controlling for school type

The lack of a clear understanding of what constitutes 'same' in 'same type of school' has prevented this methodology from being directly applied in examination comparability studies. This definition problem led to Quinlan's introduction of Delta analysis (1993). This analytical method was developed to allow for variations in school type when comparing different examination grade distributions. The method is most frequently used for investigating examination comparability between different examining groups/boards as illustrated by the GCSE Inter-Group Research Committee's study of 1995 (SEG, 1995). This analytical method only allows for gross variations in school type such as funding status in independent and state schools, and organization such as schools accommodating 11-16 and 11-18 aged students, rather than more nuanced issues related to ethos and school practices. The following example is adapted from Comparative Statistic Charts, 1975, SRAC (1976). Examination A has a larger proportion of students from independent schools than Examination B as shown in Table 3.4.

Table 3.4 Observed Examination Comparability: School Types and Delta Analysis

	Examination A		Examination B		Total	
School Type	**Entry**	**Pass**	**Entry**	**Pass**	**Entry**	**Pass**
Independent	2,000	1,200	500	400	2,500	1,600
Comprehensive	500	100	2,000	500	2,500	600
Total	2,500	1,300	2,500	900	5,000	2,200
Pass %	52%		36%		44%	

The principle of the analysis is to calculate for each school type, for example independent and comprehensive schools, an overall pass percentage i.e. the mean pass percentage for the compared examinations for each school type. In the example this is:

Independent Schools 1,600/2,500 = 64%

Comprehensive Schools 600/2,500 = 24%

These figures are taken as defining an overall standard and for each cell in Table 3.5 an expected number of passing students is calculated using this standard mean instead of the examinations' own pass rates as shown in Table 3.4. These expected numbers of passing students are then summed for each school type for each examination to give an overall *expected* pass rate for each examination, which is then compared with the *observed* pass rate. In the example, Examination A has an

observed pass rate (o) of 52% but an expected pass rate of (e) of 56%. The difference (Δ) is therefore Δ (e-o) = + 4%. For examination B Δ = 32-36 = - 4%. One might conclude that allowing for school type, Examination A is more severely graded than Examination B.

Table 3.5 Expected Examination Comparability: School Types and Delta Analysis

	Examination A		Examination B		Total	
School Type	**Entry**	**Pass**	**Entry**	**Pass**	**Entry**	**Pass**
Independent	2,000	1,280	500	320	2,500	1,600
Comprehensive	500	120	2,000	480	2,500	600
Total	2,500	1,400	2,500	800	5,000	2,200
Pass %	56%		32%		44%	

The validity of this conclusion rests in part on the basic assumption that within each school type the schools entering for a particular examination are representative of those in the type as a whole. This will not hold in reality, for example schools may be experiencing re-organization. Moreover, schools do not choose their examinations for a particular subject at random (Newbold, 1995). However, these assumptions are inherent in any analysis based only on this source of information. Another assumption made in this type of analysis is that the ratio of boys to girls within each school type does not vary from one examination to another. Again, this does not hold in reality. Thus Delta analysis is of limited value in resolving the problems associated with examination comparability. Given the foregoing discussion my view is that school variables cannot be assumed to serve as surrogates for student variables, a view supported by the work of Mortimore and Whitty (1997).

Controlling for multiple variables at multiple levels e.g. school, syllabus, examination, social and student

Theoretically this methodology above all preceding versions could validate the assumptions upon which statistical studies of examination grading standards comparability are based. As Cresswell notes in his doctoral thesis (1997) some studies (Brimer *et al*, 1978; Cresswell and Gibb, 1987) have collected data about schools and students that could be used in such an examination comparability investigation. Attempts to do so have been hindered by the complexity of needing to control for so many different variables and their interactions. Many statistical 'hurdles' have been encountered. Recent advances in multilevel modelling statistical techniques (Goldstein, 1995)

have been used. All such efforts have failed (Cresswell, 1997) because the techniques are based upon assumptions that have construct problems, such as the assumption that students interact with assessment artefacts in identical ways, as discussed earlier.

3.3 Comparability and human value judgements

The limitations of the various treatments of the variables that influence students' examination performances in technical comparability studies are widely acknowledged (Johnson and Cohen, 1983; Forrest and Shoesmith, 1985; Cresswell, 1997; Goldstein, 1995). All ignore constructions of success that might be variously assigned to different assessment domains.

In response to the limitations of the technical approach, which presumes assessment is an objective process divorced form social and cultural influences, some examination comparability studies recognise the implications of a constructivist view of mind where meanings are constructed rather than given and knower and known are inseparable. These have attempted to account for the subjective nature of the process by focusing on the importance of human judgements of value.

3.3.1 Cross moderation: catering for professional judgement

Cross moderation has been used by GCSE examining groups from the late 1980s[3] (Stobart, 1989; Ratcliffe, 1994; ULEAC, 1995; Stobart *et al*, 1994; SEG, 1995; WJEC, 1995; MEG, 1995), principally to monitor inter-group comparability of grading standards in examinations for the same subject. Subject experts (usually examiners) from the different GCSE examination groups act as scrutineers for rating the quality of work of the other groups' examination populations drawn from the grade C, A and F boundaries. The ratings given by the scrutineers indicate whether the work is judged to be better than, worse than or typical of the grade boundary region. The GCSE groups view the subjective nature of this concept as being problematic:

> *Such a concept is intrinsically hard to define. It is best, perhaps, seen in terms of what is not: work in a boundary region is work that does not clearly merit either the higher or lower grade forming that boundary.*
>
> (SEG, 1995, p. 5)

[3] This practice remains current.

Nevertheless, the method is perceived by examining groups as a valid means of them demonstrating that their students are awarded the same GCSE grade for the same level of attainment regardless of the awarding body, despite social and political concerns to the contrary (Wilmut, J. 1996). It is, however, used in conjunction with objective, technical treatments of examination performance such as a statistical analysis of examination results and an analysis by factors of the different groups' syllabuses. Any findings from the subjective cross-moderation view of comparability are still seen as requiring verification from the traditional objective technical treatments (SEG, 1995). Nevertheless it reflects an understanding that assessment validity relies on multiple sources of information (Messick, (1989).

However, it can be argued that the examining groups use cross-moderation as a technical method for resolving comparability as a technical problem. The groups hold the view that examination papers given the same professional judgement by subject experts should be graded the same by moderators. Through the process of scrutinising papers, moderators refine their professional judgment and obtain agreement about the judgement criteria they use (Houston, 1980), which are then used to cross moderate grades across the different examining groups.

This view of cross-moderation ignores the complexity of the judgement process itself. Even if a consensus view of the judgement criteria can be obtained, moderators may still differ in the way in which they apply them. They may agree about the reasons (criteria) for their judgements but differ in the values/weightings (judgements) they assign to them in their application. The criteria may be likened to Fogelin's 'agreed reasons or premises'; their application requires another set of prescriptive premises (Fogelin, 1967), for example they may agree that students need to achieve on certain types of calculations for grade A status but may differ in what they consider to be an appropriate level of correct response to these calculations.

Another problem with cross-moderation emerges when considering the issue of principal concern in the current research - examination comparability across different subjects. How might one obtain a consensus view of the criteria for examiners with differing subject specialisms? The practical problems of doing so in different subjects would certainly include finding examiners sufficiently knowledgeable to make the required value judgements in more than one subject. Theoretically it also flies in the face of Sadler's view of educational evaluation involving the use of

'tacit standards' held by a 'guild of professionals' (Sadler 1985, 1987, 1989). In the awarding process examiners of a particular subject come to an agreement about value judgements by discussing their reasons and values. From long experience of students' work and the ways in which it has been rewarded, as well as interactions with one another, teachers and examiners of a subject (guild of professionals) come to a shared understanding and acceptance of what standards should be attributed to students' work. It is therefore difficult to see how an examiner of one subject, and therefore of one particular 'guild of professionals' could possess the tacit standards of another.

3.3.2 Using a 'social value' meaning of examination comparability

Another conception of examination comparability claims to overcome the difficulties identified above. At the beginning of the research for this thesis (1995) it was only just beginning to enter assessment discourse[4]. It identifies examination awarding as an evaluative process in which equivalence or comparability is redefined not as a technical quality but as a qualitative dimension that is the value given to students' attainments by the examiners. Examination grades are taken to be human responses to students' measured attainment, rather than the attainment itself. The value judgements do not ascribe a property or properties to students' work and thus, it is claimed (Fogelin, 1967; French *et al.*, 1987; Billington, 1988), it should be possible to assign equal value (same grades) to qualitatively dissimilar attainments (different subjects/assessment domains): a requirement not fulfilled by any of the aforementioned approaches and methods.

Under this 'social value' meaning of examination comparability, the use of national examination results as the basis for selection depends upon *'a general acceptance that the judgements of examiners are valid and accurate'* (Cresswell, 1997, p. 76). Certainly the users of examinations such as students, teachers, employers and other selectors need to accept the decisions of examiners for the associated examination system to fulfil its purposes. The 'social value' meaning of examination comparability takes examinations to have comparable standards when:

---

[4] This notion of comparability of examination grading standards is developed in Cresswell's unpublished thesis of 1997.

- students for one of them receive the same grades as students for the other whose assessed attainments are given equivalent value by examiners, and;
- the examiners are accepted as competent to make such judgements by users of the examination outcomes.

Within the 'social value' meaning of comparable examination standards examiners must take account of the social value of the associated assessment domains if their judgements are to be accepted by users. The test of whether or not they are doing so derives from whether or not that acceptance is forthcoming. In other words, the tacit standards adopted by the examiners (Sadler, 1985) must reflect the views of the wider group of examination users not just their own as identified in the traditional approach to cross-moderation.

Cresswell's definition of comparability in terms of the evaluations of examiners who are accepted by users as competent assumes that the measurement of comparability requires asking users whether their acceptance is forthcoming. This would be difficult in practice, as Cresswell himself acknowledges (Cresswell, 1997, p. 81). He omits details of how it might be expedited for existing syllabuses and only speculates about the practicalities of doing so for examinations based on new syllabuses. How are users to be defined: who is to be included? How is 'acceptance' to be measured? When should such measurement take place - during the grade awarding process or afterwards?

Traditionally, the grade awarding process is conducted by examiners and examining group staff. It is only since 1994 that the Schools Curriculum and Assessment Authority (SCAA), which by 1997 changed to become the Qualifications and Curriculum Authority (QCA), appointed assessors to attend grade award meetings to exercise influence over the decisions made and, to some degree, ensure that examination standards better reflect the values of society as a whole. There are significant practical problems associated with broadening the range of participants in the grade awarding process to include a representative sample of the different types of examination users to accommodate Cresswell's conception. There is also a problem in defining what level of user acceptance supports a claim that examinations are comparable in their grade awarding standards.

Under the 'social value' conception of comparability, examiners must take into account the wider social value of the syllabuses followed by students if they are to make judgements of the value of students' achievements that are accepted by users as being appropriate. How one might determine whether this has occurred, and how any attempts to gauge users' acceptance of examiners' evaluations as being valid in terms of content or construct validity (Messick, 1988) let alone in terms of Messick's (1989) unified concept of validity, is unclear in Cresswell's thesis. Cresswell appears to assume that if a social value approach to assessment comparability is taken then validity is addressed in all its forms. Arguably, he appears to jettison validity, which in terms of Messick's unified view relates to the social mediation of the technical aspects of assessment and the social justice of the use of its technical outcomes i.e. the assessment process, outcomes and their use. Methodologically, Cresswell's conception therefore does little in the way of providing evidence of examination comparability yet implies that it is available. This I suggest is because of Cresswell's limited consideration of assessment as a process. He only applies a social mediation perspective to the meaning of grades i.e. the outcomes rather than the *process* of assessment. Aspects of the processes by which assessment outcomes are achieved, for example teachers' selection of syllabuses and student tier entry decisions, and the social mediation of students' interaction with items are not considered in Cresswell's conception. As noted earlier, this social mediation of the assessment process prior to grade awarding can undermine comparability. Cresswell's conception only appears to consider teachers as an examination user group rather than part of the process. Cobb (1999), whilst only considering influences within the classroom, identifies teachers as part of the assessment dynamic with academic success and failure in the classroom being neither a property of individual students nor of the instruction they receive but cast as a relation between individual students and the practices that they and the teacher co-construct in the course of their ongoing interactions. My view is that a similar co-construction needs to be taken into account in any consideration of academic success and failure in terms of the GCSE grade outcomes relevant to my research.

3.4 Theoretical and methodological considerations

The technical approach of using examination grade distribution comparisons is widely adopted by examining bodies and the media for the purpose of judging the comparability of examinations. The

technical response to comparability was driven initially by a psychometric model that assumes innate general intelligence, which is predictive of future achievement and independent of environmental factors. As more awareness of the influence of social factors on learning and its assessment has come into the field, the technical approach has moved to take them up as variables, as for example in the 'same school type' Delta analysis. The taking up of social factors as variables in examination comparability studies reflects a shift in a view of achievement, but the view of mind is still predominantly seen as one of the individual mediated by social factors. In calling for quantitative comparisons to be made between qualitatively differing attainments, none of the methods appear satisfactory for viewing the human mind as socially mediated, and variable and unlimited in its predisposition and preferences.

The measurement of comparability has also proved to be problematic and a source of conflict across time amongst different groups such as examination users and assessment technicists. The objections have been located in a concern for validity. One way of measuring comparability may be considered to be valid by some groups but not by others, where validity, of whatever type is deemed to be included in Messick's (1989) unified concept of its nature and refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores. Conflict arises from decisions about what constitutes firstly, evidence - that a method does not take account of this factor or that variable - and secondly, theoretical rationales, for example views of mind, for making valid interpretations and inferences from grade outcomes in order to obtain a measure of comparability. In writing about the types of evidence that may be used in determining assessment validity Messick (1989) notes that the only form of evidence bypassed or neglected in these traditional formulations of validity is that which bears on the social consequences of test interpretation and use. As Murphy and Ivinson note: '*treating assessment as a technical device isolated from social and cultural influences means that attention is not paid to those involved in the assessment process and their intentions*', (2004, p. 372).

When I came to this research it was as a physical scientist used to investigating the natural world within a quantitative paradigm. At that time I held a constructivist perspective of mind in that I viewed learning as people making sense of the world by forming mental analogies of how the

world works so that they can interpret or make sense of new information. Learning for me was not a process of absorbing information but was an active process of meaning-making. A cognitive constructivist view sees learning as personalised because people make sense of their interactions with the world and with others in unique ways. It follows, therefore, because experience is personal that the mental analogies we form of the world will differ from those of other people shaped as they are by our experiencee use our analogies as lenses through which we view new experiences and this leads us to pay attention to some features in events and phenomena and not others.

Constructivist theorising typically gives priority to individual cognition and its processes. Theorists do however differ in the way that the individual and the social are seen to interrelate. For example von Glasersfeld (1989), a radical constructivist, defines learning as self-organization highlighting an individual and local view of mind. He does, however, acknowledge that this constructive activity occurs as the individual (student) interacts with other members of the community (for example me as teacher). Bauersfeld similarly sees learning as occurring in social interaction and emphasises that *'learning is characterised by the subjective reconstruction of societal means and models through negotiation of meaning in social interaction'* (1988, p. 39). Negotiation is seen as a process of mutual adaptation in the classroom microculture in the course of which the teacher and students establish expectations for others' activity and obligations for their own activity (cf. Cobb and Bauersfeld, 1995; Voigt, 1985). As a constructivist teacher my primary concern was for my students' ways of knowing within the classroom which were developed by mutual adaptation; my students adapted their thinking in response to the learning opportunities I provided and I adapted these in response to my students' articulated understandings.

I aligned ontologically with what Guba and Lincoln (1985) call 'constructed reality'. They exemplify this with a courtroom, in which the defence and prosecution using the same events, persons and objects, construct separate realities for the jury's consideration. 'Reality' may be constructed in multiple ways from multiple rationales, and 'truth' is what is understood. Even though events, persons and objects are tangible things, the meanings and wholeness derived from or ascribed to such tangible phenomena in order to make sense of them, organise them, or reorganise a belief system, are but constructed realities. The sum of these constructed realities will

never represent the wholeness of these tangible realities. Therefore, the purpose of my research is to describe the phenomena in our experience and the relationship amongst them, not to speculate about some reality beyond that experience (Hammersely, 1995). A researcher can *'at best ... feel that he has advanced his problem along an infinite path ... there is no final accumulation and no final solution'* (Vidich and Bensman, 1968, p.396; cited in Peshkin, 1993).

A constructivist theory of learning emphasises the development of conceptual understanding where concepts are seen as *'basic units of knowledge that can be accumulated, gradually refined and combined to form ever richer cognitive structures'* (Sfard, 1998, p. 5). Sfard (1998) points out that when a constructivist view is adopted, any discussion about how people learn is underpinned by the metaphor of 'acquisition'. In this view learning is gaining possession of a commodity - knowledge, which is perceived as an entity that individuals possess and can use. Epistomologically, as at this time I saw knowledge as being constructed, it follows that the knower and the known are inseparable. Also according to Maykut and Morehouse (1994), from my ontological position as a researcher my values will become embedded in my research and thus in the way I investigate examination comparability. As events are mutually shaped, I anticipate discovering multidirectional relationships. Furthermore, my research findings will need to value context sensitivity, any explanations of my findings will be tentative and bound by time and place, for example within WJEC examination arrangements, for those particular students whose examination results I investigate, and during the time of my data collection.

The classification of research into quantitative and qualitative is based on the types of data and analyses used.

> *Data are representations of phenomena in nature, society, education and culture. Research activities are polarised into qualitative and quantitative classifications based on how phenomena are represented.*
>
> (Ercikan and Roth, 2006, p. 16)

Such polarisation is not meaningful because, as argued by Ercikan and Roth, natural and cultural phenomena are simultaneously quantitative and qualitative. The phenomenon central to my research, examination performance, has both quantitative and qualitative representations. For example it is 'quantitative' when represented by the numbers of boys and girls being allocated grades as measures of success. It is 'qualitative' when considering influences such as teachers' tier

entry decisions on girls' and boys' examination performances. As Warwick notes *'every method of data collection is only an approximation of knowledge ... all are limited when used alone'* (Warwick, 1973, p. 190). Consequently, I adopt a multiple methods approach in the thesis for the purpose of expansion, that is *'to expand the breadth and range of research* [into examination comparability] b*y using different methods for different components of inquiry'* (Johnson and Onwuegbuzie, 2004 p. 22) and with a greater consideration for the interpretability of the outcomes than currently in the public domain. Ercikan and Roth (2006) argue that full investigations of phenomena need to consider both quantitative and qualitative aspects, which they see as lying on an inference continuum with quantitative and high inference at one pole and qualitative and low inference at the opposite pole. They propose an integrated approach to educational inquiry in which the questions asked determine the modes of inquiry that are used to answer them. The questions addressed in two of my research aims include: are there relationships between students' GCSE science subject performances and are there any sex sub-group effects (research aim one); are there relationships between students' GCSE performances in the science subjects and variables such as paper construction factors (research aim two)? Due to the type of data these questions require for analysis, a mode of inquiry is sought which is quantitative. Furthermore, with my overarching interest in the thesis being to explore sources of invalidity in assessment claims, I necessarily draw on quantitative data because grade distributions are the phenomena that are at the centre of the claims I want to consider. In keeping with my epistemological position described above, I do not use technical analyses as a means of establishing truths or cause and effect relationships and so associate my quantitative study with low not high inference. Rather I consider them valuable as a means of exploring and *illuminating* potential effects and relationships that might indicate issues about sources of invalidity in relation to the comparability of subjects. In that sense and again in keeping with my epistemological position of anticipating the discovery of multidirectional relationships, I anticipate the research will reveal even more complexity than hitherto.

As the quantitative approach for research aims 1 and 2 is distinct in its methodology, type of data and analysis, I discuss this aspect of the research design and methods in this Chapter and the findings in Chapter 4.

3.5 Ethical considerations for the quantitative investigation

I was fortunate to know the Director of Research at WJEC who knew of my interests in assessment and had already consented to assist me with providing data and advice should I engage in assessment research. Examining group practices, examination data and students' performances are all potentially sensitive issues and access to previous examination comparability studies conducted by WJEC had already been denied to me for this reason. Such constraints would also serve to influence the design of this research. WJEC agreed to supply students' examination grades in computer print out form only on the understanding that the data would not be placed in the public domain in any form that would identify either individual schools or students. The same confidentiality issue applied to the other GCSE examination groups I approached to supply performance data. Having access to only hard copy from WJEC meant that a considerable amount of time was needed to create a database that would be appropriate for the use of computer statistical packages.

I made clear my research aims and strategy in my initial contacts with examining group personnel and came to an agreement with them in terms of accessing data with minimum disruption to their workings. I also kept them abreast of my analysis for fairness of interpretation. The Director of Research at WJEC agreed to assist me in selecting appropriate statistical treatments. Examining group personnel stated that they did not wish to see any sections of my thesis or related reports. I sent letters of thanks to examining group personnel on completion of my data gathering and analysis.

3.6 My quantitative research design

3.6.1 The relationships to be investigated

The first aim of this research stems from teachers' concerns with students' GCSE grades on different WJEC science examinations. By investigating the nature of students' performances across different GCSE subjects, I could begin to explore the foundation of those concerns. The traditional technical methods discussed were limited so I planned to extend the quantitative analyses to consider comparability in terms of:

(i) sub-groups i.e. boys and girls;

(ii) over time, which hitherto had not been explored for WJEC;

(iii) taking account of a wider number of variables that might influence what grade distribution is achieved than hitherto considered;

relating (iii) to considerations of the sub-group effects in (i) and trends over time in (ii).

A longitudinal study was required to explore whether any patterns in GCSE performances were sustained over time. My interest in patterns over time was in the potential to provide greater understanding of comparability issues. For example the quantitative study of the performances from one examination session might reveal differential performances between the three science subjects. If the same differential performance is sustained over succeeding examination sessions, this richer set of data provides more opportunities for understanding comparability issues.

By maintaining a quantitative approach but taking account of additional contextual variables, I could make the findings more illuminative. Given time and resource requirements, I had to limit the number of the variables for my quantitative investigation. These variables included: the nature of the students' examination centres; the coursework, its assessment and administration, and weighting in final grade allocations for GCSE biology, chemistry and physics; the nature of the cognitive demands of the associated examination papers. I limited these considerations to the WJEC populations of this research. This contextual understanding would enhance my interpretation of the quantitative findings for illuminating the notion of 'gradeness'. Stobart *et al.* (1992) attempted some analysis of this kind in their study of the gender differences in performance in GCSE English and mathematics, based on the Assessment of Performance Unit question descriptors. I decided to use the findings (Appendix 2) from my previous study on the cognitive demands of examination papers (Benson, 1995) to inform the interpretation of the outcomes of the quantitative analyses and the methods used are described in section 3.6.7. If the nature of the coursework, its assessment and administration, and weighting in final grade allocations had changed across the examinations investigated in this research, I would be less secure in making valid comparisons of the examination performances across time. It was not my intention to make a detailed study of the part that coursework contributes to overall GCSE science grades, rather to consider the potential influence of the variable, coursework. My intention was to identify whether there had been any disparities in the nature of the coursework and its administration between the three science subjects and across the years of the study that might in turn impact on my

populations' achieved grades. The following aspects of coursework were considered: whether it was practical based; the general types of activity expected of students; whether it was teacher-assessed; the percentage of the total marks allocated to it (weighting); possible changes in the aforementioned.

Institutional factors rather than factors relating to the assessments themselves are also shown by the literature to be influential. For example, type of examination centre has not been previously considered in any published GCSE science grade comparability study, although it was acknowledged that this factor could influence GCSE subject grade distributions (SEG, 1995). Examination centre is taken to mean any educational institution that prepares and enters students for the GCSE examinations considered in this study. I decided to identify the nature of my examination populations in terms of aspects of their examination centres. My intention was to understand how similar were the study's WJEC populations in respect of their centres' nature across the years of my study, as changes in the nature of the populations in these respects might influence their GCSE performances. For example, traditionally, independent schools generally achieve better GCSE grades than their comprehensive counterparts (TES, 25.08.95; *TES*, 26.08.05). The co-educational and single sex nature of an examination centre has also been identified as an important factor in public science examination / science assessment performances (Bell, 1989). Delta analysis is a statistical treatment that aims to control for centre type influences and validate comparisons of examination grades of candidatures from different types of centres / schools. As noted earlier (see 3.2.2) the assumptions upon which this method is based are flawed. My view is that at best I can know what types of centres my examination populations are based in so as to qualify my examination comparability findings. For example, if I find that say, my 1993 and 1994 populations are respectively skewed towards independent and comprehensive type schools, I can use these findings to interpret and qualify the outcomes of comparing these populations' achieved GCSE grades. The aspects of the centres' nature that I explored are described with the methods used in section 3.6.7 to avoid repetition here.

A consideration of the relationships between science, English and mathematics achieved GCSE grades was considered useful for investigating the nature of students' performance on GCSE

science examination papers. With this approach to the database, I identified the following specific investigations for the research:

- the relationships between students' performances in WJEC biology, chemistry and physics GCSE examinations;
- the relationships between students' performances in WJEC biology, chemistry, physics GCSE examinations and their average GCSE grade scores;
- the relationships between students' performances in WJEC biology, chemistry and physics GCSE examinations, their average GCSE grade scores and their English and mathematics GCSE performances;
- the relationships between students' sex (where this term refers to the student's sex, and not gender as a social construct) and their achieved WJEC GCSE biology, chemistry, physics, English, mathematics grades and average GCSE grade scores;
- a comparison of the aforementioned relationships identified for WJEC with those of other examination groups to explore, in a limited way, the transferability of the WJEC findings.

The literature review revealed the practice of using a reference test to obtain a measure of students' general attainment or subject specific attainment (Bardell *et al.*, 1978) against which students' attainment in a particular subject can be compared. However, because of the issues of test bias and theoretical incompatibility between the reference test and the examinations let alone the practical difficulties of my obtaining consent to approach Year 11 students to administer a test, I decided not to adopt a reference test method. GCSE group personnel instead used students' average GCSE grade scores as a measure of general attainment and a basis for comparing achievements on different GCSE subjects. Scores of 0 to 8 were allocated respectively to GCSE grades A* to U. For every candidate their average GCSE grade score was found by dividing the sum of their subject scores by the total number of their obtained subject scores. WJEC routinely calculated students' average GCSE grade scores and were willing to provide them for me. I discussed the challenges to this method of comparing students' achievements on different GCSE subjects earlier. However, examining groups argue (WJEC, 1999) that by taking large numbers of students there is more chance of a similar spread of affective factors such as motivation across the subjects. This makes the method more tenable. However, there may still be differences in the

number of science subjects that the students have taken and I address this issue below when identifying my student populations. Given the availability of the data and these caveats I decided to use average GCSE grade scores as a means of quantitatively investigating performances on the different science GCSE subjects for large samples of students.

3.6.2 Database parameters: strategic considerations

Time-related Issues

Due to grade appeals' procedures, access to examination data would only be possible after the end of September in each examination year. Given that I was in employment throughout the lifetime of this research, it was anticipated that the data sets would not be analysed until the Spring following the previous year's examination session. The 1993 examination session was the first source of data made available to me by WJEC. If data similar to that collected for the 1993 examination session were also collected for the 1994 session, I would be able to look for sustainability in GCSE performance patterns. During the initial stages of this study there was indecision regarding the reporting of GCSE achievements in terms of the ten levels (later reduced to eight plus an extension level) associated with the National Curriculum for students aged 11-16. However, during 1994, the requirement for GCSE examining groups to report on the ten levels was rescinded and grades A* to U were used (this applies in 2008) in contrast to the 1993 arrangement of grades A to U. The 1995 examinations were the first to reflect the introduction of the National Curriculum at GCSE. Furthermore, due to the Code of Practice (SCAA, 1995b), from the 1995 examination session, students had to be entered for all three of their separate science GCSE subjects (biology, chemistry and physics) from the same syllabus suite, for example Nuffield science syllabuses, administered by the same examination group. This restriction did not apply to independent schools; as discussed this exemplifies the influences that reduce the validity of comparing examination performances across schools. New syllabuses for GCSE biology, chemistry and physics were also examined for the first time in 1995, again as a result of the Code of Practice (ibid.). Thus, 1995 separate science GCSE examinations would be associated with factors such as changing entry patterns for examining groups and compared with previous years, differences in scientific content within examination papers. I needed to take these changes into account.

Consequently, it had to be recognised that in this field of research the parameters within which students' achievements would be measured and reported on at GCSE were subject to change. Nevertheless, conducting the same type of investigation of comparability of students' examination performances for more than one examination session, say for each year of 1993, 1994, 1995, might reveal (i) the existence or absence of patterns in GCSE performances over time and (ii) changes in methods of measurement and reporting of students' achievements. Even though those findings would be subject to extensive qualification in their interpretation, such a strategy would enrich the research findings, and provide increased opportunities for understanding the meaning of comparability. For this reason a repeated measures strategy (Coolican, 1994) was adopted, in the sense that the methods used to study the 1993 WJEC GCSE examinations would also be applied to the 1994 and 1995 examination sessions. The rational for this decision was that:

- taking consecutive years of examination sessions would minimise the changes in methods of measuring and reporting students' GCSE achievements and thus minimise some of the factors that would reduce the validity of conducting a comparability study of examination performances;
- taking three years worth of examination sessions would facilitate a comparison over time of students' GCSE performances to identify any sustained patterns in subject comparability;
- three years would be within the time available for my collection of examination performance data whilst living near to WJEC offices and my contact there being in post.

Variety in Sources of GCSE Performance Data

Examining groups other than WJEC were approached with a view to providing GCSE data for the examination sessions 1993 to 1995. This was to enable me to explore the relationships between the combined boys' and girls', and separate boys' and girls' achieved biology, chemistry, physics, English, mathematics grades and average GCSE grade scores *across* examining groups. Of these, only the University of Oxford Delegacy of Local Examinations (UODLE) gave permission to access their data, and in this case it was for Southern Examining Groups' (SEG) GCSE science examination data (UODLE, which did not itself directly offer GCSE, was affiliated to SEG which merged with NEAB and AEB in 2000 to form AQA, the Assessment and Qualifications Alliance). This offer was taken up with an assurance that data sets would be provided on disc saving time in

data set creation. However, by early 1995, despite many communications with UODLE personnel, receipt of the data was still awaited. I approached SEG directly in the summer and assurance was given: the data would be forthcoming but by now, it would not be possible to provide 1993 GCSE examination data. Later that same year the required data for the examination sessions 1994 and 1995 was received. Consequently, the findings from the WJEC study could only be explored in relation to another examining group for the last two years of the main WJEC study.

3.6.3 Identifying an appropriate student population

The grade awarding process is claimed to bring students' performances to a common denominator so that one can argue that a grade A on one examination is equivalent to a grade A on another examination (Wilmott, 1994). At the time of my data gathering GCSE groups were not required to provide 'profile reporting' by Attainment Targets. This means that the components (biology, chemistry and physics) in GCSE Double and /or Single Award Science were not individually graded - they may have existed in a notional form, but were not formally stated nor for all grade boundaries (Newbold, 1995) (in 2008 grades for biology, chemistry and physics components exist for co-ordinated science but not for integrated science syllabuses). If they had, it would have been of help in considering the relative performances in the separate sciences within Double and Single Award Science. Without that grading, there would be little value in comparing marks on the components (ibid.). Consequently, I decided to focus only on the three sciences as separate GCSE subjects in Triple Award GCSE Science.

This was an appropriate decision in terms of the associated student population. At the time of my research, schools in England, not uncommonly, entered their students for different separate science subjects with different GCSE examination groups (Newbold, 1994, personal communication). This situation makes for incomplete data that would be problematic for my selecting separate science GCSE as the basis of my study. All state schools in Wales are registered WJEC examination centres. Consequently the vast majority of young people in Wales are prepared for WJEC's GCSE examinations and this avoids problems of incomplete data.

I also decided that it would be more informative to compare the performances of those students who had taken *all three* of the separate GCSE science subjects than compare the performances of the whole examination populations (not the same groups of students) for each of

GCSE biology, chemistry and physics. In the latter case, comparing the achieved means for biology, chemistry and physics candidatures would not really address the issue of whether students' are more severely graded in one subject than another. Therefore, my focus was only on those students who had attempted three science subjects in Triple Award GCSE Science.

Students are entered for their GCSE biology, chemistry and physics in terms of particular tiers. For example, in the 1995 examination session, the separate science subjects were offered in three tiers, with only one giving access to the top three grades (A*, A and B). I considered it

on those students who had attempted an option in each science subject that gave access to the same grade range. Otherwise, comparisons would be made between a students' grades on say, biology examination papers offering access to grades A*-C, with chemistry examination papers only offering access to a maximum of grade C. Such disparities in students' separate science tier entries are not uncommon (WJEC, 1994, personal communication) because teachers perceive their students as having differential abilities in the separate science subjects and select tiers for their students to match those perceptions.

This entry factor, that to an extent prejudges the performance of students, was seen to reduce the validity of this comparability study. I sought to explore assessment artefacts between papers across subjects. I did not wish to add to that factors which might arise between papers within a subject as previous subject-pair comparisons had done but without due consideration for validity issues (NEAB, 1993; UCLES, 1994). By focusing only on those students who had attempted GCSE biology, chemistry and physics in a tier giving access to the same grade range, this potential source of invalidity was taken into account as far as was possible. No other published GCSE science comparability study had controlled for this variable. The tier chosen was the one that offered students access to the top four GCSE grades. This tier was chosen as in my experience, and confirmed by NEAB and UCLES personnel (ibid.), it is generally only those students who are considered capable of the top GCSE grades that are entered for all three separate science subjects. Furthermore, students entered for any tier of WJEC's GCSE separate sciences come largely from state rather than independent schools unlike the situation with some English GCSE examining groups at the time of this research (Newbold, 1995, personal correspondence). In selecting the tier giving access to the top GCSE grades in the separate science GCSE examinations for my

comparability study, the associated student population would remain representative of the WJEC GCSE examination population as a whole in terms of examination centres. The 1995 examination session saw a different allocation of awarded grades to tiers than for the two previous years. Consequently to cover the same grade range as for the 1993 and 1994 examination sessions, students were considered who had attempted tier 03 (option R) with its expected grades B-A*, in addition to and separately from those students who had attempted tier 02 (option Q) with expected grades B-E.

During the grade awarding process, students whose scores correspond to grades other than those allocated to their entry tier range may be allocated an exceptional grade to allow for 'mistakes' in tier entry. For example, if a student entered for tier 02 offering grades B-E achieves an examination score equivalent to that required for grade A they may be awarded a grade A as an exceptional case. Similarly, a student entered for tier 02 might only achieve a score just below that required for grade E and so be awarded grade F as an exceptional grade. For tier 03 a grade C might also be awarded to students whose achieved scores do not equate to the grade range A*-B. Consequently, when dealing with and interpreting the grade outcomes of the study's WJEC 1995 examination sessions the existence of exceptional grades needed to be considered. The other WJEC and SEG examination sessions i.e. 1993 and 1994 did not have overlapping grade tiers and thus, exceptional grade awarding did not occur.

Therefore, the WJEC population that was identified included students who had attempted all three of the WJEC separate GCSE subjects, biology, chemistry and physics, in the tier giving them access to the top GCSE grades, namely tier 02 (grades A-U) in 1993 and tier 02 (grades A*-U) in 1994. For 1995, students who had attempted these GCSE science subjects were considered in two separate groups, namely those who had sat tier 02 in all of these science subjects and then for the same subjects, those who had sat all tier 03s. When looking across examining groups the SEG examination population was selected to match as closely as possible that for WJEC in the terms outlined above. The number of students within these identified WJEC and SEG populations for the examination sessions ranged from a minimum of 387 to a maximum of 1761.

3.6.4 Data gathering: identifying the required data and its method of collection

Students are traditionally allocated examination numbers in examining groups' databases. Students' schools/places of examination are also traditionally allocated examination centre numbers rather than using actual names. These centre and student numbers were used as a means of identifying the students in this study's populations and served as the main references on the GCSE computer printouts from WJEC and computer discs from SEG.

The students in tier 02 in examination sessions 1993-95 and additionally tier 03 in the 1995 examination session had to be identified from the printouts of the achieved grades for the *whole examination populations* for each of the GCSE subjects, biology, chemistry and physics. To identify these students physics was chosen as it had a smaller number of students than those for biology and chemistry. The students identified in the whole of the physics' examination population were then searched for in the whole examination populations of GCSE biology and chemistry. Those students who were found to have taken the physics 02 tier and tier 02 in each of biology and chemistry were selected and their GCSE student numbers and examination centre numbers identified. Students who had taken tier 03 in each of these science subjects in 1995 were identified in a similar manner.

The sex of students taking GCSE examinations is traditionally shown alongside their examination numbers on WJEC computer printouts. However, the average GCSE grade score for each student is not routinely shown on such printouts and had to be requested. WJEC generated a computer printout of students' GCSE subjects, their associated grades and average grade scores for the *whole* of the GCSE chemistry candidature. Thus, once the study's WJEC population had been identified from cross-referencing the whole examination population lists, all necessary data was available.

WJEC also provided the English and mathematics GCSE grades for the study's populations. A number of students' were shown as not having been entered for GCSE English and/or mathematics in the same examination session as the majority of their other GCSE subjects. Initially it was assumed that these students had been entered for GCSE English or mathematics in the year previous to the majority of their GCSE subjects i.e. fast tracked able students. Towards the end of this study's quantitative data collection (1996), efforts were made to trace these students'

English and mathematics grades. The WJEC 1992 -1995 examination databases were checked to see if the students had:

(1) achieved a WJEC GCSE grade in these subjects a year earlier or a year later than the year in which they had sat the majority of their subjects in the same examination centre;

(2) achieved a GCSE grade in these subjects in an examination centre different to that associated with the majority their GCSE subjects;

(3) entered for these GCSE subjects with another examining group.

The vast majority of the students' English and mathematics GCSE grades were eventually identified. The numbers in each of the above categories (1) to (3) varied with the examination session (1993-95). Consequently, the subsequent analysis of the data involving GCSE English and mathematics was dealt with in two ways for each year of the WJEC study:

(1) the data for students' *WJEC* GCSE biology, chemistry, physics, English and mathematics grades obtained in the *same* examination session formed one data set for subsequent analysis.

(2) the data for students' GCSE biology, chemistry, physics, English and mathematics grades, where the latter two subjects were obtained in *any* examination session and with *any* examination group, formed another data set for similar subsequent analysis involving English and mathematics GCSE grades.

For each year of the study, the vast majority of the students in the study's population fell into category (1).

SEG provided students' examination numbers, examination centre numbers, sex, GCSE biology, chemistry and physics achieved grades and average GCSE grade scores for those students who had sat the 'extended' option in 1994 and the 'high' option in 1995 in each of the three science GCSE subjects. The SEG population for the study had proportionally greater numbers of students than the WJEC population. Given the difficulties encountered in obtaining complete data sets of English and mathematics GCSE grades for the WJEC population, I judged it impracticable to examine the SEG population's GCSE English and mathematics grades. Consequently, my intention to explore the transferability of the WJEC findings in other GCSE examining groups was amended to compare the relationships between combined boys and girls, and separate boys' and

girls' achieved biology, chemistry, physics and average GCSE grade scores from SEG with those for WJEC.

3.6.5 Data processing

SPSS for Windows was used to create a database and to analyse the data for each examination session. For each WJEC examination session data was entered under field names as shown in Table 3.6.

Table 3.6 Field Names in the WJEC Database

Field Name	Variable / Data entered
centre	Examination centre code
cand	Individual student's examination code
sex	1.00 for a boy and 2.00 for a girl
avgrade	The average grade score shown on the computer printout
eng	The student's GCSE grade for English (if known)
maths	The student's GCSE grade for mathematics (if known)
bio	The student's GCSE grade for biology
chem	The student's GCSE grade for chemistry
phys	The student's GCSE grade for physics

In preparation for the analysis stage of this study and in accordance with technical examination research practice (Wilmott, 1994; Kelly, 1976; Nuttall *et al*, 1974), all of the fields containing GCSE grades were recoded from string to numeric form. The original and changed values are shown in Table 3.7.

Table 3.7 Changing GCSE grades into a numeric form within the WJEC database

1993 Examination Session

GCSE Grade	A	B	C	D	E	F	G	U	No grade
New Value	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00

1994 and 1995 Examination Sessions

GCSE Grade	A*	A	B	C	D	E	F	G	U	No grade
New Value	0.00	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00

Statisticians might argue that allocating a scale of 1 to 8 for examination grades A to U is inappropriate because the scale 1 to 8 is an interval scale and the differences in the marks between the grades are not the same. For example: grade A might be allocated to all marks above 84 out of 100; grade B is then deemed to be appropriate for all marks between 75 and 83 i.e. within a band of 9 marks; grade C might then be allocated to all marks lying between 65 and 71 i.e. within a band of 7 marks. Based on my examining experience I anticipated that the grades would not be equal integers. Technically I could not address this. The use of interval scales in grade analysis has been used by educational researchers including the Schools Council and NFER at least since the 1960s. Furthermore, Wilmott (1994) used calculations of standard deviations based on this common practice of allocating grades to an interval scale and found this to provide results which allowed comparison with other studies. Thus, I decided to adopt the same practice of examination grade coding. To facilitate cross-referencing, the value labels of 1.00 for 'boy' and 2.00 for 'girl' were entered in the field / variable, sex. Each examination session's data, and in the case of 1995, each of the tier 02 and tier 03 population's data, was used to create separate data sets for ease of reference within the database. All of the subsequent analysis involving WJEC's GCSE results was conducted on these data sets.

Many more variables than those requested were included in the SEG data files, for example student's date of birth and GCSE grades for all attempted subjects. Consequently, all variables not required for the study were collectively put into a separate field. Two SEG data sets were processed, one for the 1994 data and one for the 1995 data. The value labels used in the WJEC data sets were used to produce data sets that could be subjected to similar analysis. Finally, the WJEC and SEG data sets were copied to Excel files to allow graphical representations of the data analysis.

3.6.6 Choosing analytical procedures

The relationships between students' performances in WJEC biology, chemistry and physics GCSE examinations.

GCSE examination groups routinely conduct subject-pair analysis (Nuttal *et al.*, 1974; NEAB, 1994; UCLES, 1994) of achieved GCSE grades as a means of gauging the relationships between students' performances in different subjects. In this method the GCSE grade obtained by a student in one subject is compared with the GCSE grade the same student obtains in another subject.

Differences in the grades for one student are explained away, for example different levels of well-being on the days of the subjects' examinations. However, the subject-pair method is based on the premise that if the grades differ significantly across *all* of the students sitting these examinations, then the subjects are not of equal 'difficulty'.

The subject-pair method of analysis does not represent a definitive statement on the relationships between students' examination performances and by implication, the relative severity/leniency of examination subjects. It merely offers an estimate (Nuttall *et al.*, 1974) and even then is conditional on the compared examinations having similar standard deviations. However, because the subject-pair method continues to be used routinely by GCSE examination groups (UCLES, 2001; UCLES Group, 2005), it was decided to adopt it (given the standard deviation requirement was met), as to do so, would allow comparison of my research findings with others.

The unbiased mean total (UBMT) method referred to in a discussion of Nuttall's (Nuttall *et al.*, 1974) work relies on the same assumptions as the subject-pair method. It was decided it was an inappropriate method to use because it does not use a contribution from the subject under consideration, a fact that serves to reduce the validity and reliability of its findings (ibid.). It is not widely used in studies relevant to this research and therefore would not enable comparisons. The iterative method first used by Kelly (1976) and used under the name of 'correction factor approach' by Fitz-Gibbon and Vincent (1994) in their analysis of GCE 'A' level grades, was also considered. The iterative method is used to analyse a full matrix of students and subjects, with four subjects and iterations being required to produce reasonable levels of accuracy (Kelly, 1976). The associated calculations are complex and Fitz-Gibbon and Vincent in their analysis of GCE 'A' level results (ALIS), use a special software package created by their statistician colleagues at the University of Newcastle. Attempts to obtain a copy of this software were unsuccessful and because the study needed a method that would deal with just three subjects (biology, chemistry and physics), it was decided not to adopt the iterative method.

Because I sought to investigate the relationships between students' performances in GCSE subjects, it was necessary to identify associations and differences in those performances. Relationships, and in particular associations, are studied statistically by means of correlation, which

may be defined as, '*the measurement of the extent to which pairs of related values on two variables tend to change together*' (Coolican, 1994). To examine relationships between students' performances there are three variables, the biology, chemistry and physics GCSE grades. The strength of the relationship between any two variables is expressed on a scale ranging from -1 (perfect negative) through zero (no association) to +1 (perfect positive). The figure that expresses this type of relationship is the correlation coefficient of which there are several types. The two types of correlation coefficient in common use are Pearson's product-moment (r), generally used with interval or ratio data as a parametric test, and Spearman's rho (r_S), which may be used on non-interval data in a non-parametric test. Pearson's r requires data to be normally distributed and is appropriate for application to large sample sizes, for example data collected from at least one hundred people (Abouserie, 1992). Conversely, Spearman's r_S does not make the assumption of normally distributed data and may be used on relatively small sized samples.

This study's data certainly fulfilled the requirement of large sample sizes, 631 students being the smallest sample (the number of students in 1993's dataset). The data had also been processed to an interval scale, thus again indicating that Pearson's r would be appropriate. However, the criterion of normal distribution of data was not securely fulfilled (see Figure 4.1 in Chapter 4) as expected in an educational assessment system. All of the examination performance distributions were positively skewed - some extremely so, some approximated normal distribution, for example the SEG 1994 dataset, but several did not. The lack of normal distribution was a significant factor, the type of data being of less importance as Spearman's r_S may be used on data other than ordinal (The Open University, 1996). Consequently, in consultation with a statistician, it was decided to apply Spearman's r_S to this study's data in the calculation of correlation coefficients and subsequent tests for inference.

However, as a check on the revealed significance of the calculated correlation coefficients, the 1993 dataset (with its deviation from normal deviation and positive skewness) was used to obtain Pearson r values. There appeared to be little difference in the calculated Pearson r and Spearman r_S values and significance testing findings (Table 3.8): at most the Pearson r and Spearman r_S values only differ in their second decimal place.

Table 3.8

Pearson r and Spearman r_S Correlation Coefficient Values using WJEC 1993 Data

		Chemistry	Physics
Biology	Pearson r	0.5348	0.5783
	Spearman r_S	0.4957	0.5723
Chemistry	Pearson r		0.6171
	Spearman r_S		0.6010

Generally parametric procedures (here Pearson r) are often markedly more powerful than their non-parametric counterparts (here Spearman r_S). That is, generally a parametric procedure will more frequently reject a null hypothesis than will a non-parametric test designed to perform the same function. This can be attributed to the parametric procedures using more of the available information, such as the deviations from the mean of the scores in the analysis. Non-parametric procedures more frequently rely upon frequency count and ranking procedures, thus discarding some of the information available in the data. Thus there are power differences in the parametric and non-parametric techniques. For example, if a difference is highly significant by using a non-parametric test, then it has a very high probability of being significant when using parametric statistics (Popham and Sirotnik, 1973). However, the reverse is not true. Consequently, as a high level of significance was found with a relatively weak statistical test for the 1993 data set (here, Spearman r_S and its associated significance testing), one might expect a very high probability of also finding significance when using Pearson r.

Thus Spearman r_S was considered to be the most appropriate treatment for this study's data. As for Pearson's r, the higher the value of r_S, the more positive the correlation (as the value of one variable increases, the value of the other variable also increases); the lower the value (below zero) the more negative the correlation (as the value of one variable increases, the value of the other variable decreases). Spearman's r_S values for pairs of data within the subjects' data sets, for example the biology grades with the chemistry grades, and the biology with physics and physics with chemistry grade data sets, provide a measure of the association of the students' performances in these subjects.

Correlation studies provide measures of association between variables, which in this study are taken to be the grades for the different GCSE subjects. A positive association indicates that

students who obtain a high grade in one subject are likely to obtain a high grade in another subject – but not necessarily the same grade. On the other hand, the statistical term, agreement, measures the number of occasions that students obtain the same grade in two different GCSE subjects. As noted earlier, some science teachers perceive students as being more likely to achieve a grade A in one science subject than in another: in this sense they have a notion of 'subject-specific gradeness'. This notion can be explored by measuring agreement between students' awarded GCSE grades in different science subjects. Since there would be some agreement if the students were allocated GCSE grades at random, a kappa treatment is used to allow for this chance agreement, and can be thought of as the proportion of cases agreeing after allowing for chance agreement. Hildebrand, Laing and Rosenthal (1970) give an account of how kappa may be calculated.

A kappa treatment requires there to be the same grade range in both of the considered GCSE subjects i.e. if grades A, B, C and D have been awarded in one subject, then they must all have been awarded and with no other grades in the second subject. This requirement was expected to hold for at least the majority of the awardable grades in the study's datasets because of the large student numbers. Nevertheless, it was a requirement that needed to be checked for compliance in each of the paired GCSE subjects of the study. Where this requirement was found not to hold, then kappa could not be calculated and these datasets needed to be considered for adjustment. The first step in the adjustment involved frequency tests on all of the study's science subject (biology, chemistry, physics) data sets for one particular examination session. Then each of these subject frequencies was examined to see which grades were not common to all the subjects. The students with these particular grades were then identified and omitted from all of the considered science subjects' data sets. If the number of omitted students was small compared with the total number of students in the data sets, this data adjustment seemed reasonable and did not detract from comparing correlation coefficients and kappa values for paired subjects. However, how small is a 'small' number of students needed to be considered for each of the study's data sets and a value judgement taken of the merits of still calculating kappa.

Kappa values may be placed on a scale of agreement (Landis and Koch, 1977) as follows:

0.00	poor	0.41-0.60	moderate
0.01-0.20	slight	0.61-0.80	substantial
0.21-0.40	fair	0.81-1.00	almost perfect

It is theoretically possible to have a high positive correlation between the grades in two subjects but at the same time, low agreement (Bell, 1999). A correlation coefficient of value, 0.7, and a kappa value of 0.2 could be obtained from a comparison of the same pair of subjects' data sets (ibid.). By calculating correlation coefficients and kappa values more detail about the nature of the relationship between the awarded grades in different subjects could be provided.

Any identified associations, agreements and lack of agreements in the students' GCSE performances might be usefully explained by conducting a descriptive statistical treatment of their achieved GCSE grades. The mean and standard deviation of the processed GCSE grades for each science subject's examination session data set were considered to be the most useful descriptors to facilitate such an explanation. Bar charts are used to present the calculated frequencies in an easily assimilated form.

The relationships between students' WJEC science performances and average GCSE grade scores

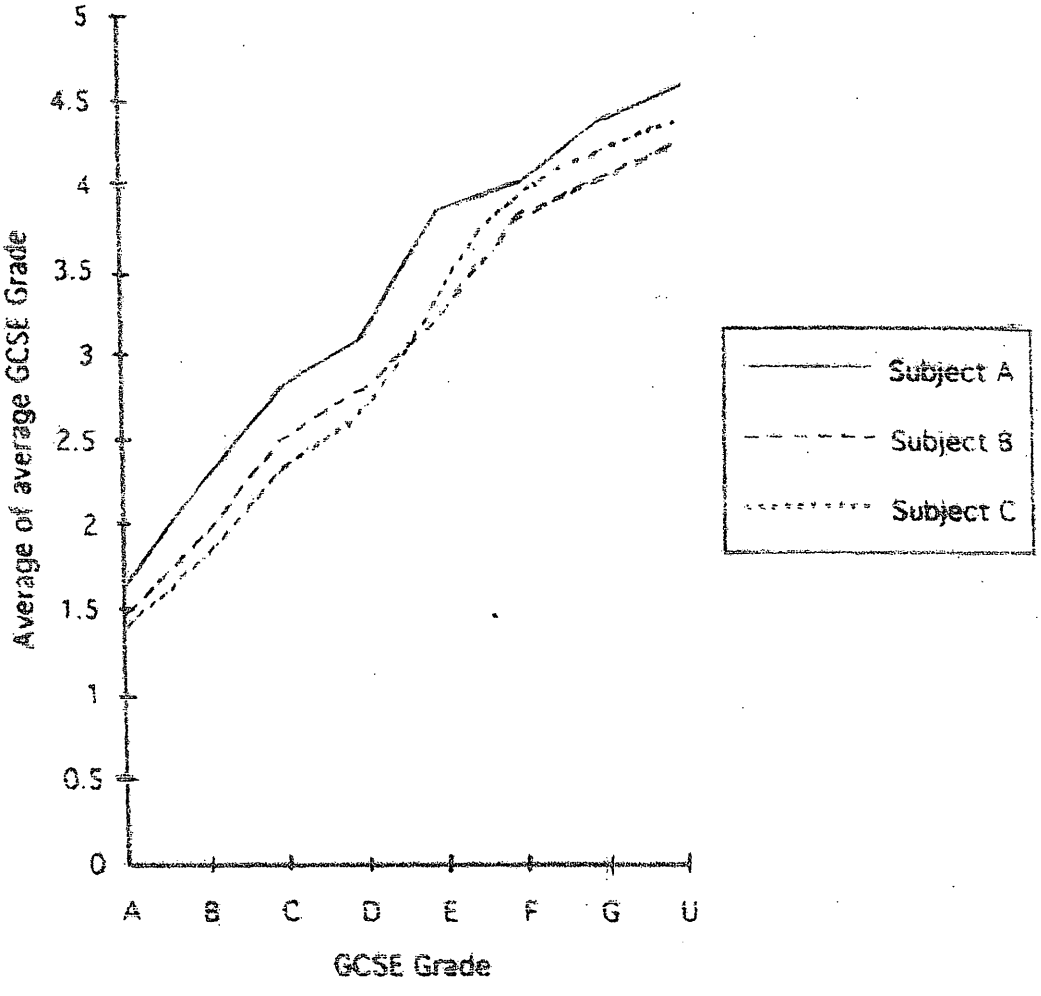
This investigation extended my study of comparability. It was prompted by the advice received from examining group research personnel (particularly those of WJEC) that students' average GCSE grade scores are useful in providing a standard against which GCSE subject performances can be compared. Consequently, it also seemed appropriate to use their recommended associated analytical procedures. Each subject (biology, chemistry and physics) was considered in turn. The students achieving a particular grade in the subject had the average of their average GCSE grade scores calculated for each GCSE grade in the subject. A graph was then plotted of the average of the average GCSE grade scores against the achievable GCSE grades (see Figure 3.1). This procedure was repeated for all of the science GCSE subjects. Within the resulting graphs, the shallower the slope of the line and the further down the y axis a subject's line appears, the more severely graded or 'harder' the subject appears to be (WJEC, 1994), and is generally interpreted in this way.

The relationships between students' WJEC science performances, average GCSE grade scores and English and mathematics GCSE performances

There were difficulties associated with obtaining the English and mathematics GCSE grades for all of the students in the study's WJEC populations. For those students whose biology, chemistry, physics, English and mathematics GCSE grades and average GCSE grade scores for a particular

Figure 3.1

Graph of Mean GCSE Grade against GCSE Grades for Subjects A, B and C.



examination session were known, the association of these variables was investigated by a correlation study with the calculation of Spearman's r_S correlation coefficients and subsequent testing for statistical significance.

Measuring the agreement between students' awarded GCSE grades in the different science subjects and (i) English and (ii) mathematics and then between English and mathematics was also conducted to extend the study's exploration of the notion of 'gradeness' being stable across subjects', that grades having common currency across subjects. The subject datasets were adjusted in the manner discussed earlier but with English and mathematics awarded grades also being considered so that the same types of grades are present for *all* pairs of compared subjects. This ensures that the same students are compared in each of the subsequent kappa calculations (Table 3.9).

The percentage adjustment in data sets for WJEC 1995 tier 03 and 02 was so large as to prevent legitimate comparisons between correlation and kappa values for the BCP pairings without substantial qualification. It could be argued that correlation and kappa could be calculated on the kappa data sets. However, by using the larger data sets for correlation coefficient calculation, the more reliable are the subsequent correlation findings for particular subject pairings. I was also primarily interested in measuring association and agreement rather than in comparing the two. For these reasons I decided to use the separate data sets as shown in Table 3.9.

Table 3.9 Data set adjustments for kappa calculations

Number of Candidates		**Before adjustment**	**After adjustment**		**%Change**	
			BCP	BCPEM	BCP	BCPEM
WJEC 1993		631	618	588	2.0	7.3
WJEC 1994		792	787	778	0.6	1.8
WJEC 1995	Tier 03	387	387	336	0.0	13.2
	Tier 02	608	578	494	4.9	18.8
SEG 1994		1001	998	...	0.3	...
SEG 1995		1761	1758	...	0.2	...

Note:

BCP = biology, chemistry and physics grade matched datasets.

BCPEM = biology, chemistry, physics, English and mathematics matched grades datasets.

... denotes English and mathematics grades are not available for creating these matched grade datasets.

To help understand any identified associations and reveal possible differences in the students' English and mathematics GCSE performances the means and standard deviations of the students' GCSE performances for each of the study's subjects were calculated, these being the minimum descriptive statistics required. Cross-tabulation was applied to students' available paired English and mathematics processed GCSE grades to examine underlying differences in students' performance.

The relationships between sex group and achieved WJEC science, English, mathematics performances and average GCSE grade scores

The analysis of comparability next considered the sub-group effects i.e. the performances of the boys and girls. The research is primarily interested in identifying any differences between these performances. Consequently, the boys' and girls' GCSE performances were explored using two types of analysis. The significance of any sex-related performance differences was examined using *t* - tests in an unrelated design, for which the conditions of use are: a need to look for differences rather than correlations; interval or ratio type data; data needing to be drawn from two independent groups; data needing to satisfy parametric assumptions. All of these conditions were met, again the robustness of the *t* - test accommodating some deviation from normality. The two independent groups of data were the boys' and girls' achieved GCSE processed grades and average GCSE grade scores. Any identified significant differences were then explained by descriptive statistics i.e. each sub-group's mean and standard deviation values in the associated subjects' and average GCSE grade scores.

In addition, I decided to provide bar charts showing the two sub-group's relative performances in the subjects of biology, chemistry, physics, English and mathematics to make the study's findings explicit. First, SPSS was used to run a cross-tabulation of the variable, sex, with each of the variables: biology, chemistry, physics, English and mathematics processed GCSE grades in the form of frequencies for each grade category. Such analysis can only consider overall sub-group differences. To consider what this might suggest about 'gradeness' and to allow a more trustworthy analysis of overall differences, analysis was based on percentages of *each sub-group's population* rather than the study's particular examination session's whole population. This is in line

with Bell's recommendation from his study of sex differences in performance in Double Award Science GCSE (Bell, 1997).

Comparing the relationships identified for WJEC with those of SEG.

To facilitate such a comparison, it was considered necessary to subject the data from the other participating examination group, SEG, to the same treatments as those used for the WJEC part of this study, with the exception of the English and mathematics considerations as explained earlier. The SEG data was 'eyeballed' (Coolican, 1994, p251) to establish that it met the normality and homogeneity assumptions for parametric treatments, although as stated earlier, the selected treatments are very tolerant of deviations from these assumptions. Indeed, the SEG data deviated less from these two assumptions (see tables in Appendix 3) than did the WJEC data, probably because larger data sets were involved, 1001 and 1761 being the respective number of students in the 1994 and 1995 SEG data sets.

3.6.7 Setting up the investigation of the relationship between students' GCSE performances in different science subjects and associated variables

As discussed in section 3.6.1 my intention was to understand the *context* of my technical analysis. I am using context here in relation to the examination variables within which my data sets emerged.

Exploring the nature of the centres

To do this I first set up examination centre identity profiles for each of my WJEC examination populations (see Appendix 3) by obtaining the five digit number called an identity code, which WJEC allocates to each examination centre, for each student in my populations. These centre identity codes with their respective school / college name and address were obtained directly from the paper examination datasets that WJEC had given me for their science examination results. I calculated how many of my students in my populations came from the same centres and how these numbers and the lists of the centres changed across my study's 1993, 1994 and 1995 populations. By 'eyeballing' (Coolican, 1994, p251) the centres' addresses I could also see whether there were any clear changes in geographical location of the centres across the years of my study.

The WJEC grade result computer printouts decribed each student's sex and centre in relation to three number codes. These codes identified: the status and type of the centre, for example secondary, selective, tertiary college; the centre's nature of control or government, for example maintained, independent; the age range of students, for example 11-16, 16-19. Co-

educational or single sex organisation was not identified. As this aspect of a centre had already been identified as an important factor in public science examination / science assessment performances (Bell, 1989), I judged it necessary for interpreting the study's findings. This additional information was collected by means of telephone contact.

It was necessary to further process the WJEC datasets to analyse entry patterns in terms of sex and centre type. The WJEC data sets were amended to include four new fields as shown in Table 3.10.

Table 3.10 Field Names for Centre Types in the WJEC Database

Field Name	Variable / Data entered
status	centre status and type (numbers 1-9), for example secondary, selective, tertiary college.
nature	nature of control or government (numbers 1-8), for example maintained, independent.
age	age range of students (numbers 1-9), for example 11-16, 16-19.
censex	sex of students present in the school (numbers 1-3)

Value labels of 1-9, 1-8 and 1-9 were respectively allocated to the fields 'status', 'nature' and 'age' in accordance with the numerical codes outlined in WJEC's documentation (WJEC, 1994). Single sex or co-educational was entered as a new field 'censex' with value labels of 1 for a centre with only boys, 2 if only girls were present and 3 if the centre was co-educational. As a result, all of the information necessary for subsequent analysis of centres and their types was present on the amended WJEC data sets. Offering both the number of centre categories and the percentage of the population's students present in the categories for each of the fields, status / nature / age / censex, provides a profile of the student populations for each of the research examination sessions (1993-95). This provides contextual information for interpreting the relationships between the students' GCSE performances and takes some account of the caveats raised regarding the technical approach to comparability as discussed.

Exploring the nature of the coursework

It was not my intention to make a detailed study of the part that coursework contributes to overall GCSE science grades. I wished to identify whether there had been any disparities in the nature of

the coursework and its administration between the three science subjects and across the years of the study that might in turn impact on my populations' achieved grades. To obtain this information I analysed the syllabuses (WJEC, 1993, 1994, 1995) by 'eyeballing' (Coolican, 1994, p. 251) them in terms of: whether the coursework was practical based; the general types of activity expected of students' engaged in coursework; whether the coursework was teacher assessed; the percentage of the total marks allocated to the coursework (weighting); and, possible changes in these aspects during the examination sessions (1993, 1994, 1995).

Exploring the cognitive demands of the examination papers

At the time the research began (1995), the WJEC syllabuses for GCSE biology, chemistry and physics framed their assessment objectives in terms of knowledge / recall, understanding, application, analysis, synthesis, evaluation and experimental / practical work. The latter was addressed through coursework. Prompted by teachers' concerns through 1993 – 1995 about comparability of students' WJEC GCSE biology, chemistry and physics grade outcomes, I had already conducted an analysis of these 1993 - 1995 science papers for these types of cognitive demands prior to commencing the research for this thesis.

I had obtained an expert consensus view of the cognitive demands of the examination papers associated with the current study i.e. a face validity measure (Coolican, 1994) of this construct from my post graduate secondary science student teachers nearing completion of their professional training. Of these experts the biology graduates viewed the biology papers, the chemists the chemistry papers and the physicists the physics papers so that the science of the various questions would be understood and they could focus on the type of cognitive demand of individual questions.

I grouped the cognitive skills of knowledge / recall, understanding, application, analysis, synthesis, evaluation into three groups, namely, knowledge / recall; comprehension and application; and analysis, synthesis and evaluation for simplicity. The students allocated each examination question to one of the three groups based on their view of the predominant skill demand. I did not train the students, it was their individual views that I sought from which I could then obtain a consensus view. How the data was collected, processed and analysed may be found in the ensuing paper (Benson, 1995). The percentages of total marks allocated to each of the three

groupings of cognitive demands are shown in Appendix 2 and used in Chapter 4 as a part of the interpretation of the patterns in students' GCSE science performances.

3.7 Way forward

The outcomes of the statistical analyses of examination performance detailed above are typically used by examining groups to explore comparability in terms of difficulty. However this fails to take account of the many influences already discussed that might cause differences in examination performances which may have little to do with inherent difficulty. In the next chapter the findings of my analyses addressing research aims 1 and 2 are discussed in terms of what they indicate about comparability and the sources of influences to explore with teachers.

CHAPTER 4

Exploring 'gradeness': the quantitative analysis

In assuming comparability examining groups continue to use some of the methods described in Chapter 3 to correct for deviations or to indicate for example that standards are falling in respect of school teaching, students' skills or subject difficulty. My interest is in the meaning of 'gradeness' – that students' achieved grades have common currency. I intended to use some of the methods used by examining group personnel but with the strengthening of certain of their aspects, for example sampling, and with a particular interpretation taking account of the caveats that I raised about the methods in Chapter 3. I do not assume comparability as do examining groups. I explored quantitative relationships to see what 'gradeness' meant for my particular populations and in so doing, wished to expose the dynamic nature of assessment that as I have indicated in Chapter 3, is not controllable or monitorable by technical means. As I am using the methods of assessment technicists in examining groups, I also tend to use their discourse, for example severity of grading, when in fact, as I have argued in Chapter 3, differences in groups of students' examination performances may have nothing to do with severity of grading, the examinations may just be *different*.

4.1 Comparing the study's Welsh Joint Examining Consortium (WJEC) populations

4.1.1 The populations' examination centres

The examination centre profile (see Appendix 3) revealed that the number of centres increased from 1993 (53) with the addition of new centres for 1994 (64). There was an increase in the student population sizes too (1993 = 631; 1994 = 792). This might be explained by the 1994 examinations being the first in which students were required to be examined in all three science subjects, either by Double Award, Single Award or Triple Award (separate sciences) GCSE examinations. In these respects the 1994 population differs from that of 1993 in the science education background of its students and geographical distribution of its centres and in turn, whatever influences these might have on students' performances. The 1994 and 1995 profiles showed the study's populations increasing in

student numbers from 1994 to 1995. These profiles showed a greater degree of similarity than dissimilarity in centre geographical distribution, although there was a small decrease in centres based in South and West Glamorgan (centres beginning with 687) from 1994 to 1995. I can only speculate about the reasons for the decrease in West and South Glamorgan centres. Centres may have changed to another examining group or changed their policy from entering students for WJEC's Triple Award GCSE to WJEC Double Award Science GCSE. In 1995 the requirement for centres to enter their students for all three separate science subjects with the same examining group was introduced. This may have prompted some of the 1994 centres to move from WJEC for the 1995 examinations, especially if they had been accustomed to picking particular examining groups for each of the science subjects because they perceived them to favour students' attainment. It would be reasonable to assume that the introduction of new syllabuses would serve as a stimulus for centres reviewing their entry policies including whether to continue to enter their students for Triple Award GCSE or Double or Single Award Science GCSE with WJEC. These sorts of issues I intended to explore when I engaged with teachers in centres as a means of understanding the historical influences on examination comparability (Chapter 6). Appendix 3, however, shows there was a substantial core of common centres associated with each of this study's examination sessions (1993 – 1995), with an influx of new centres in the 1994 session.

For each student in each of my study's populations (see Appendix 4) I had identified their examination centre's:

(1) status, for example independent, comprehensive, college of further education;

(2) locus of control, for example voluntary aided;

(3) age range of students;

(4) intake of boys and girls,

and calculated the percentage of each population coming from all boys, all girls, and coeducational type centres.

Even allowing for the influx of new centres in 1994, Appendix 4 shows a striking consistency in the centres' status and type for all of the study's populations. More than 85 per cent of each of the

study's populations consist of students attending secondary comprehensive schools, secondary independent schools being the next most common category but at a significantly less represented level. The vast majority of the students in each population are based in maintained schools, the next most common category being independent schools. In this respect I can say that any skew towards high performance for any of my populations is unlikely to be due to a disproportionate number of independent centres (see Chapter 3.2.2). The vast majority of the students in each population are based in schools covering the age range 11-19 years with percentages ranging from 70.9 in 1993 to 82.4 in 1995, tier 03. The next most common category is 11-16 schools with percentages varying from 5.9 (1995, tier 02) to 18.7 (1993). This represents a substantial variation in the relative proportion of students in 11-16 schools across the study's populations. Nevertheless, each of the populations is dominated by students in centres that have staff and the resources to teach post GCSE, for example 'A' level. There was a striking consistency in the percentage of students who come from all boys, all girls and coeducational centres in the study's populations.

My view was that there was sufficient similarity in the nature of the centres associated with the examination sessions of this study to support the usefulness of my comparison of their students' achieved grades across time (1993 – 1995).

4.1.2 Coursework arrangements

My intention was to identify whether there had been any disparities in the nature of the coursework and its administration between the three science subjects and across the years of the study that might in turn impact on my populations' achieved grades. The aspects of coursework considered were:

(1) whether it was practical based;

(2) the general types of activity expected of students;

(3) whether it was teacher-assessed;

(4) the percentage of the total marks allocated to it (weighting);

(5) possible changes in the above (1) – (4).

As noted in Chapter 2, coursework weighting changed over the years specifically increasing from 20 per cent to 25 per cent for all of the science subjects for the 1995 assessments to reflect the

National Curriculum at GCSE. The nature of the coursework also changed at this time. For both the 1993 and 1994 examination sessions biology and chemistry coursework consisted of practical exercises written by teachers with exemplar materials to serve as guidance within similar assessment objective frameworks. They were administered and also marked by teachers. Physics had a different arrangement: teachers were expected to assess their students in a similar way to biology and chemistry teachers but the outcomes only counted as 10 per cent of the coursework weighting – the remaining 10 percent resulted from a WJEC set practical test, administered and marked by teachers. The 1995 examination session saw a rationalization of coursework assessment, with GCSE biology, chemistry and physics all requiring the assessment of students' skills to plan and carry out investigations using teacher selected, implemented and marked practical tasks. For each of the examination sessions of this study, moderation of practical coursework (including the physics practical tests of 1993 and 1994) by WJEC personnel aimed to achieve a measure of consistency in standards within and between the different GCSE science subjects across time.

Although the potential consequences of the physics' practical tests of 1993 and 1994 for students' motivation and overall relative achievements are recognized, they are currently unquantifiable. Similarly the change in weighting and nature of practical coursework first observed in the 1995 examination session might also have influenced students' motivation and achievements. Certainly the increase in weighting should have an impact on the profile of expected achievements at any particular grade boundary. It could be argued that moderation of coursework by WJEC should remove variations in standards so that grade A in 1993 is achieved with the same standard of scientific skills, including those in the practical domain, as in 1995. One could also argue that these identified differences in the nature and weighting of coursework, impact on the validity of comparing students' achievements across the different GCSE science subjects and across the study's examination sessions. I bear these arguments in mind when I discuss emerging patterns in my findings at the end of the Chapter.

4.2 Presentation of the findings

Students' GCSE results from the three WJEC and two SEG consecutive examination sessions constituted a large database. I decided to present the analysis and interpretation of findings about the

nature of students' performances in different science subjects first followed by the findings about the nature of the relationship between these performances and variables such as the examination paper cognitive skill demands. This would allow an understanding of the data to be developed and help make sense of any relationships emerging.

Each of the relationships explored is dealt with in turn. For each analytical treatment used to explore the potential relationship the data and analysis are first discussed *within* years and then *across* years to explore any consistency in findings when this approach does not introduce unnecessary repetition. Otherwise, the data and analysis from the different years are discussed together for each analytical treatment. The 1995 WJEC examination session saw a different allocation of awarded grades than for 1994 and 1993. This meant I had to consider students entered for Tier 03 and Tier 02 examination papers as separate groups. The 1995 WJEC data and analysis are therefore presented as separate Tier 03 and Tier 02 outcomes. The SEG data and analysis are presented after that for WJEC. The findings are represented in bar charts and line graphs with different colours for biology, chemistry, physics, English and mathematics. A dark shade and a hatched pattern in the subject's colour are used to represent boys' and girls' performances respectively.

Given my sample sizes and that my study is concerned with social science and educational norms for statistical significance, I have adopted a five per cent significance level for rejecting the null hypothesis as recommended by Coolican (1994). Differences or relationships in the examination performance findings are counted as significant when $p \leq 0.05$.

4.3 Exploring relationships between students' performances in WJEC biology, chemistry and physics GCSE examinations

4.3.1 Subject-pair analysis and findings

Tables 4.1 and 4.2 show respectively the mean grades and the subject pair results.

Table 4.1 Biology, chemistry and physics means – WJEC

	Subject	Mean	
1993 (*N=631*)			
	Biology	2.40	
	Chemistry	1.73	Expected Grades A-U apply
	Physics	2.48	Allocated values are A=1 to U=8
1994 (*N=792*)			
	Biology	1.81	
	Chemistry	1.49	Expected Grades A*-U apply
	Physics	1.54	Allocated values are A*=0 to U=8
1995 (Tier 03)			
(*N= 387*)	Biology	0.89	
	Chemistry	1.09	Expected Grades A*-B apply
	Physics	0.68	Allocated values are A*=0 to B=2
1995 (Tier 02)			
(*N= 610* biology and chemistry, *608* for physics)			
	Biology	2.84	
	Chemistry	2.74	Expected Grades B-U apply
	Physics	2.42	Allocated values are B=2 to U=8

Table 4.2 Subject-pair analysis – WJEC

	Mean grade (A)		Mean grade (B)		Difference (A-B)
1993	Biology	2.40	Chemistry	1.73	0.67
	Biology	2.40	Physics	2.48	- 0.08
	Chemistry	1.73	Physics	2.48	- 0.75
1994	Biology	1.81	Chemistry	1.49	0.32
	Biology	1.81	Physics	1.54	0.27
	Chemistry	1.49	Physics	1.54	- 0.05
1995 Tier 03					
	Biology	0.89	Chemistry	1.09	- 0.20
	Biology	0.89	Physics	0.68	0.21
	Chemistry	1.09	Physics	0.68	0.41
1995 Tier 02					
	Biology	2.84	Chemistry	2.74	0.10
	Biology	2.84	Physics	2.42	0.42
	Chemistry	2.74	Physics	2.42	0.32

Examining groups regard negative differences as indicating (B) is more severely graded than (A)

Given the reverse nature of the grade / score conversion where 1 = grade A (the highest awarded grade in 1993) and 8 = grade U (the lowest awarded grade), Table 4.1 shows that for the 1993 population under scrutiny, the students' mean grade is highest in chemistry, then biology followed by physics.

The subject-pair method is used by examining groups to produce estimates of subject difficulty or severity of marking. There is an assumption of comparability and deviations from this are explained from a number of premises. How deviations are interpreted will depend on what is seen to influence students' interactions with assessment items. I use the expressions of the examining groups in discussing the findings in that differences in grade achievement are interpreted as differences in difficulty of the particular examination process investigated rather than the subject per se. I use the analyses to consider differences in the grades achieved by the populations and what patterns within these there may be and what this might suggest about the meaning of gradeness taken as a common currency with a given and understood meaning. From the 1993 examination dataset in Table 4.2 examination groups interpret the findings as: chemistry is less severely graded than biology; physics is more severely graded than biology and even more severely graded than chemistry. This analysis suggests there is a tendency for students to perform at a higher level in chemistry, followed by biology and then physics.

For the study's 1994 population, where grade A* = 0 (the highest awarded grade in 1994 GCSE examinations), the students' mean grade is highest in chemistry, then physics followed by biology. The subject pair values could be interpreted by examining groups as indicating chemistry is less severely graded than physics, biology is more severely graded than physics and even more severely graded than chemistry. Therefore, the 1993 and the 1994 populations under scrutiny both appear to perform highest in chemistry, but differ in the subject that they perform least well in. The students' mean grades are highest in physics for each of the study's populations entered for Tier 03 and Tier 02 (as for 1994, Grade A* = 0). Physics therefore appears to be the least severely graded of the three science subjects for these two populations. However, these populations differ in terms of which science subject is most severely graded, this being chemistry for Tier 03 and biology for Tier 02 respectively.

The subject pair findings could be interpreted as showing a change in severity of grading across the years of my study with the most significant change occurring for chemistry becoming more severely graded and physics becoming less severely graded. The change is also most marked for the 1995 population and arguably this may be due to underlying influences emanating from the GCSE syllabus changes reflecting the National Curriculum first examined in 1995. However, one could also argue that the populations are different and they may have interacted with the assessment artefacts in differing ways to produce the subject pair findings. The subjects may not have been differently graded in severity and 'gradeness' may not really vary across subjects and time - the subject pair method findings may have masked a variety of underlying influences.

4.3.2 Correlation

The analysis is summarized in Table 4.3.

Table 4.3 Spearman Correlation Coefficients Between Biology, Chemistry and Physics Grades – WJEC

	Chemistry	Physics
Biology		
1993	0.50	0.57
1994	0.56	0.58
1995(03)	0.55	0.39
1995(02)	0.54	0.47
Mean	0.54	0.50
Chemistry		
1993		0.60
1994		0.62
1995(03)		0.50
1995(02)		0.55
Mean		0.57

N for 1993 = 631; 1994 = 792; 1995(03) = 387; 1995(02) = 610 for biology and chemistry pairings, 608 for physics pairings.

All correlation coefficient values are significant at the 0.1% level

Overall, the mean values indicate that chemistry and physics are the most positively correlated pairing followed by chemistry and biology, and then biology and physics. For both the 1993 and 1994 populations under scrutiny, there is a greater tendency for each population's physics and chemistry

grades to be positively correlated than either their chemistry and biology grades or physics and biology grades. This pattern holds for the 1995 Tier 02 population. There is also a greater tendency for both the 1995 Tier 03 and Tier 02 populations' biology and physics grades to be least positively correlated than any of the other subject pairings for these populations. One interpretation is that the physical science examinations are similar in their paper constructions and place similar demands on students in one way or another – similarities that appear to a lesser degree on the chemistry : biology and physics : biology paired examinations.

These correlation findings reinforce the notion of population interaction with assessment artefacts and so question the assumption that this can be ignored in the traditional technical approach to comparability. Arguably, if the physics and chemistry assessment domains require similar ways of thinking then one would *expect* a strong positive correlation in assessment outcomes. This could be an issue for further exploration, for example consideration of my correlation findings against my examination paper analysis.

Other patterns in correlation are subsumed within that obtained from a comparison of the mean values. The correlation coefficient values for both the 1993 and 1994 populations are in the same order of most positive for the chemistry and physics pairings, then biology and physics and least positive for chemistry and biology. The associated coefficient values for the subject pairings in these two years are also similar. This is not the case for the values for each of the 1995 Tier 03 and 02 populations, which show more dissimilarity with each other as well as with the values for the 1993 and 1994 populations. Arguably my findings indicate a change in examination outcomes occurring from 1995. Certainly new syllabuses reflecting the National Curriculum were examined for the first time in 1995. Whatever has occurred at this time, my findings suggest that there may be an impact on comparing examination performances and that it is useful to look for changes in my examination paper analysis findings between 1994 and 1995 to illuminate this further.

4.3.3 Kappa

Table 4.4 shows all the WJEC study's kappa values based on the matched grade datasets. The kappa values provide a measure of agreement between students' awarded grades in the different science

subjects, for example the extent to which it is likely that students with grade B in biology will attain grade B in chemistry whilst allowing for chance agreement. The mean kappa values indicate that there is 'fair' (Landis and Koch, 1977) agreement between students' attained grades for all of the science subject pairings. All of the study's populations, bar that of 1993, have their lowest kappa value for the biology and physics pairing.

Table 4.4 Kappa Values for Biology, Chemistry and Physics – WJEC

	Chemistry	Physics
Biology		
1993	0.136	0.232
1994	0.199	0.168
1995(03)	0.254	0.222
1995(02)	0.280	0.171
Mean	0.217	0.195
Chemistry		
1993		0.135
1994		0.271
1995(03)		0.227
1995(02)		0.282
Mean		0.229

All Kappa values are significant to 0.1%

Number of cases 1993 = 618; 1994 = 787; 1995(03) = 387; 1995(02) = 578.

The highest levels of agreement (but only still within the 'fair' range of Landis and Koch's scale of agreement) are shown most often (1994 and 1995(02)) for the chemistry and physics pairing. However, this particular subject pairing also produces the lowest kappa values in the other two examination populations (1993 and 1995(03)). No other trends in kappa values are apparent. The notion of 'subject-specific gradeness' in terms of it being more likely for students to achieve the same grade in physics and chemistry than any other science subject pairing is supported by the findings but not strongly so. Several arguments flow from this finding. Arguably, if the physics and chemistry assessments call for similar ways of thinking, something intervenes, perhaps in the marking and grade awarding processes, to reduce the possibility of identical grades being obtained by students on these two subjects. One could also say that although the correlation findings show that students attaining high grades in physics are significantly likely to also attain high grades in chemistry, arguably there

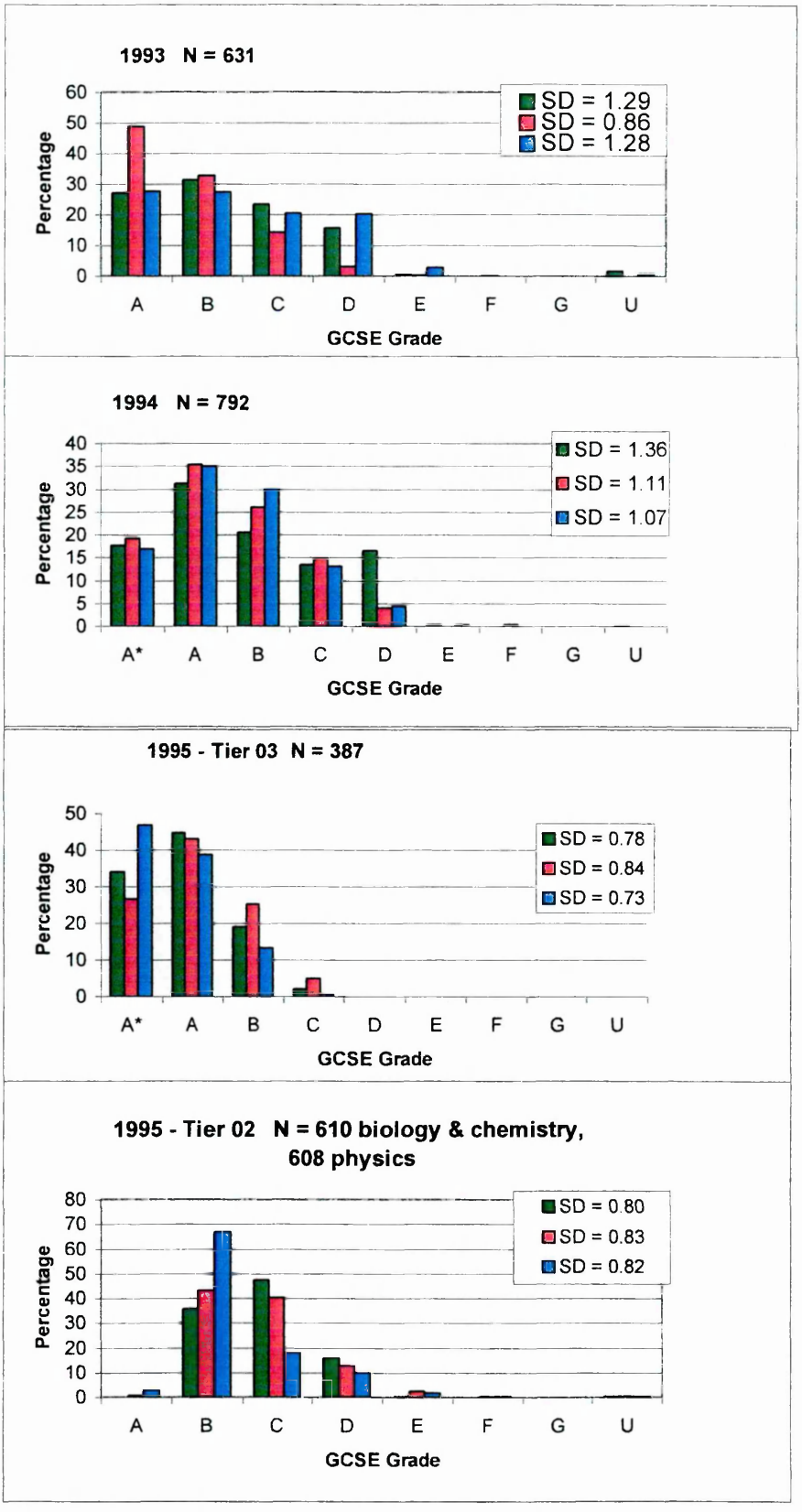
are still sufficient differences in how the students interact with the assessment artefacts to prevent them obtaining identical grades in these subjects. A composite of underlying influences could exist to result in only 'fair' kappa values being obtained. Comparing the correlation and kappa values illustrates again the complexity of examination comparability and the potential for population interaction with assessment artefacts – interactions that are largely ignored in the technical approach to comparability. Given the anecdotal evidence in Chapter 1, the finding that it is only 'fairly' likely that students will be awarded the same grade in their science subjects indicates it could be explored with teachers.

4.3.4 Descriptive statistics

Appendix 5 shows the frequency, percentage and cumulative percentage of the students achieving each GCSE grade within biology, chemistry and physics for the WJEC 1993, 1994, 1995 (Tier 03) and (Tier 02) datasets. The means and standard deviations are also shown. This information is presented in bar charts in Figure 4.1.

The 1993 standard deviation values are similar for biology (SD 1.29) and physics (SD 1.28) with students' chemistry grades being much less widely distributed (SD 0.86). Although all of the achieved science subject grades are positively skewed for the 1993 population, chemistry shows the most positive skewness, then biology followed by physics. Proportionally more students achieved the higher grades of A and B in chemistry than in either physics or biology and proportionally more students achieved the lower grades of D and E in physics than in either biology or chemistry. In terms of comparability, one could argue that it was easier to get a high grade in chemistry than either physics or biology – but this could only be said for this population, which I have selected from the whole Triple Award Science GCSE candidature for having taken all of their Triple Science GCSE examination papers in the same tier. Proportionally more students achieved the higher grades of A*, A and B in chemistry and physics (80.9 and 82.2 per cent respectively) than in biology (69.6 per cent) in the 1994 population. Furthermore, proportionally more students achieved the lower grades of C and D in biology (30.0 per cent) than in either chemistry or physics (18.8 and 17.5 per cent). Indeed, overall, there is a marked similarity in the distribution of this population's achieved chemistry and

Figure 4.1 WJEC GCSE Biology, Chemistry and Physics Grade Distributions



physics grades. Examining groups might argue that this is because they have comparable severity of grading. Unlike the 1993 analysis outcomes, the 1994 population's standard deviation values are similar for chemistry (SD 1.11) and physics (SD 1.07) with students' biology grades being slightly more widely distributed (SD 1.36). If comparability is described in terms of similar distributions of grades for the different science subjects, one could argue that these 1994 findings suggest a higher standard of comparability across the different science subjects than those for 1993. Furthermore population analyses such as standard deviations and individual analyses such as kappa significantly alter how 'gradeness' is understood. The issue of what is used to describe or define comparability for examination performances in different subjects is therefore significant. Is it the means, standard deviation values, kappa values or some other descriptor – and is what is used by one person or group acceptable as valid for its purpose by another person / group?

This was an issue that I identified for exploration with teachers to illuminate the meaning of comparability. I wanted to explore how teachers talked about examination comparability – what descriptor(s) do *they* use in their language and do they use quantitative descriptors? If standard deviation is used as a descriptor of comparability as examining groups do (WJEC, 1995; SEG, 1995), then the 1995 Tier (03) and Tier (02) show a shift towards similar values and arguably a move towards more similar severity of grading for biology, chemistry and physics than in the 1994 and 1993 data. Again I am drawn to qualify this statement as in previous analyses - it only applies to my populations. If these findings indicate such a 'change' in examination performance outcomes from 1995, are they despite of or the result of new syllabuses being examined for the first time in 1995? I have no answer but teachers' views of the impact of examining the new syllabuses might illuminate this issue of comparability.

4.4 Exploring relationships between students' WJEC science performances and their average GCSE grade scores

4.4.1 Graphical analysis

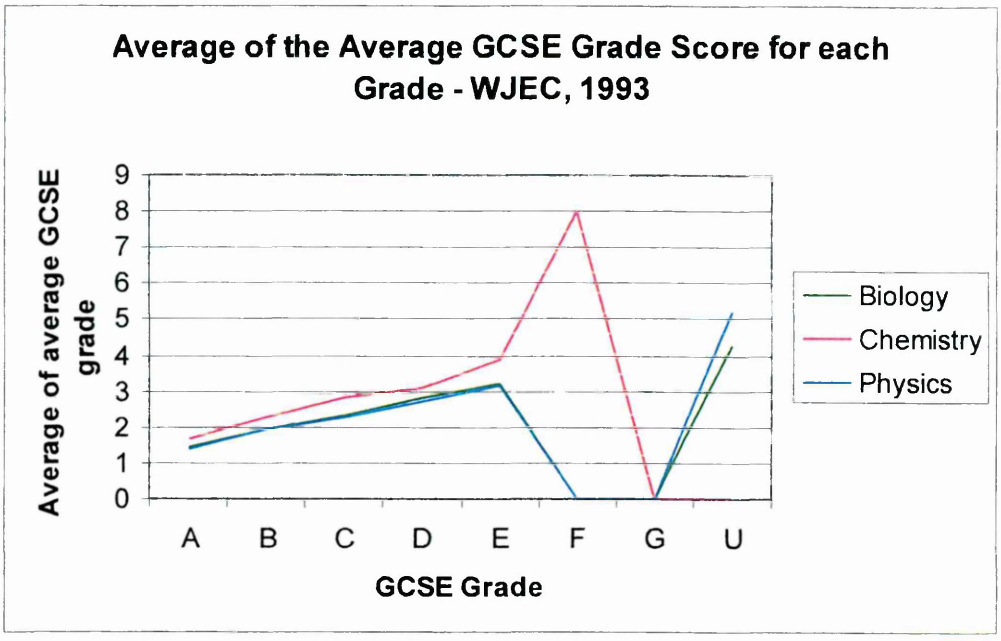
In this analysis, students achieving a particular grade in a science subject e.g. biology, have the average of their average GCSE grade score for all of their GCSE subjects calculated and this is plotted against their achieved grades in biology, chemistry and physics as shown in Figures 4.2. The gradients of the lines are used by examining group personnel to give an indication of the

relative 'difficulty' of the subjects, biology, chemistry and physics for the populations: the steeper the gradient and / or the higher up the y axis the line appears, the 'easier' (less severely graded) the subject is seen to be.

Figure 4.2 for the 1993 population shows that within the GCSE grade range A-E: the chemistry line has a steeper gradient than either of the biology and physics lines; the biology and physics lines are very similar in their slopes with a slight tendency for the physics line to be less steep than that for biology. There is a slight divergence occurring between grades B-E caused by a slight increase in the slope of the chemistry line. Within the grade range A-E, chemistry appears to be less severely graded ('less difficult') than either physics or biology, which in turn appear to be very similar in their degree of severity of grading across this grade range when compared with students' average GCSE grade scores. These findings concur with those from the means and subject pair method. Within the grade range E-U there is no apparent relationship in the slopes of the lines due to no students obtaining grade F in biology and physics, grade G in all three subjects and grade U in chemistry

For the 1994 population Figure 4.2 shows that within the GCSE grade range A*-D the biology, chemistry and physics lines are similar in their slopes with a slight tendency for the biology line to be increasingly less steep than those for chemistry and physics over the grade range B-D. Within the grade range A*-D, there is a tendency for chemistry and physics to be slightly less severely graded than biology, which in turn appears to become progressively more severely graded than these two science subjects from grades A to D. The level of 'difficulty' for chemistry and physics remains very similar across the grade range A*-D. As for the 1993 findings these for 1994 concur with their mean and subject –pair counterparts. Within the grade range E-U there is no overall pattern in the slopes of the lines due to an absence of students obtaining grade E in chemistry, grade F in biology and physics, grade G in all three subjects and grade U in chemistry and physics. The disparity in the existence of awarded grades below the level of grade D in the three science subjects indicates no relationship within the grade range D-U.

Figure 4.2 Graphical Analysis: WJEC 1993 and 1994



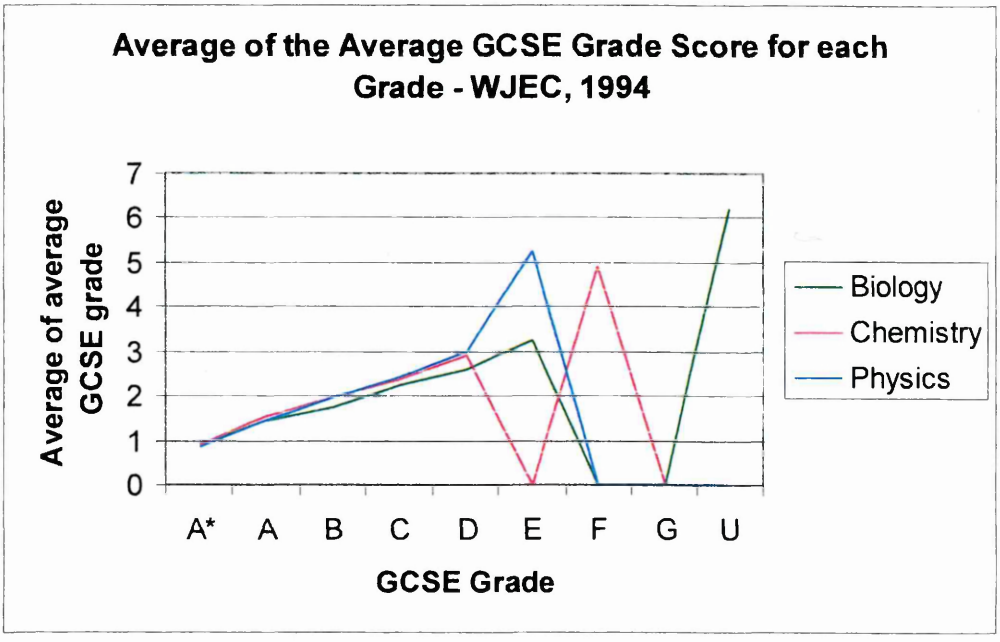
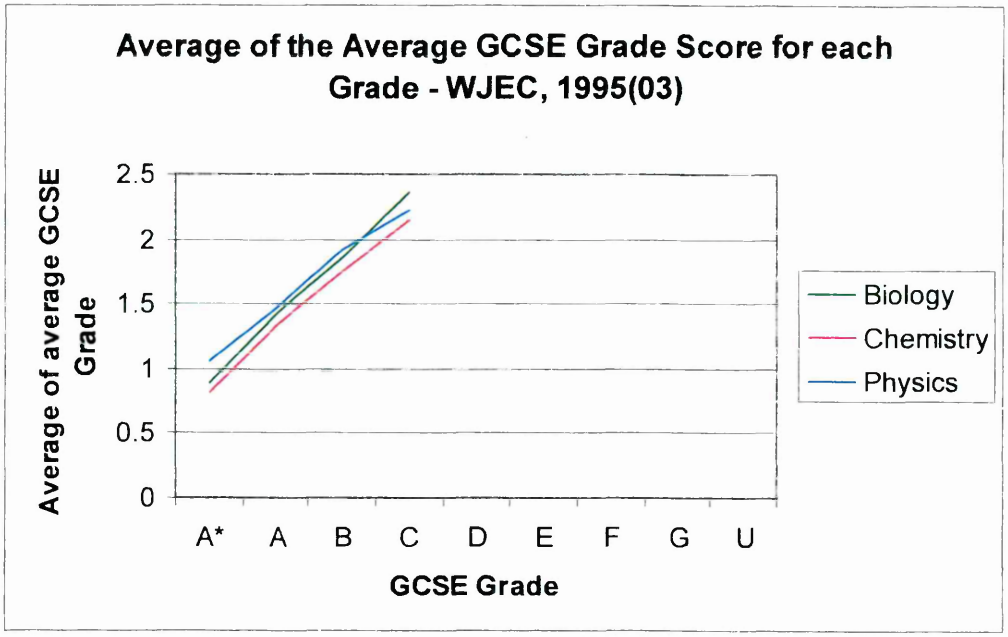
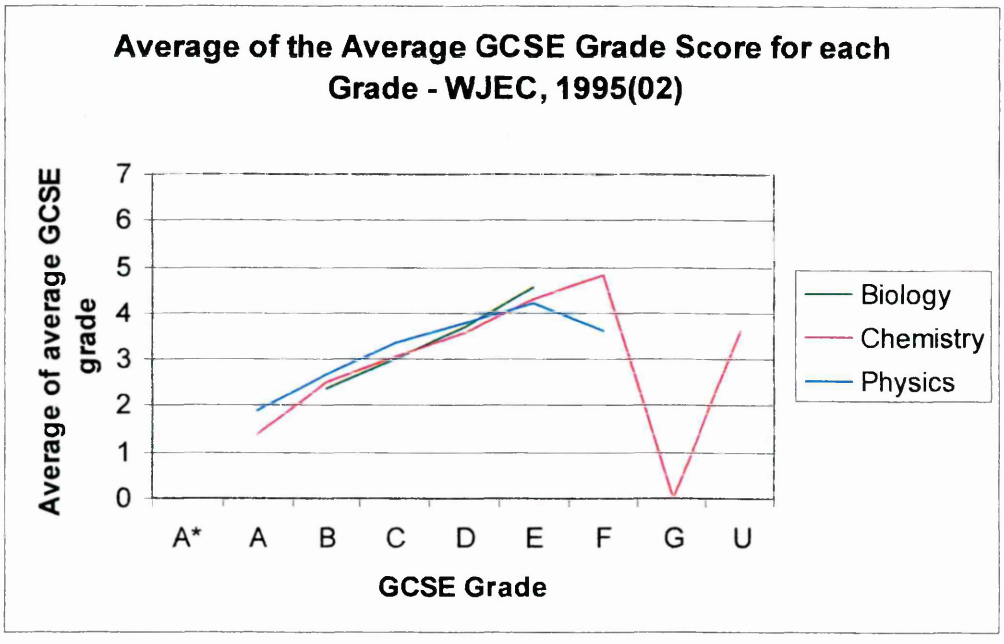


Figure 4.2 Graphical Analysis: WJEC 1995(03) and 1995(02)





The grades available for the 1995 (Tier 03) population are restricted and Figure 4.2 shows that across the GCSE grade range A* - B, the physics line lies above those of biology and chemistry, and the biology line moves from being nearer the chemistry line at grade A* to being nearer the physics line at grade A. Within the grade range A*-C, there is a tendency for chemistry to be the most severely graded, which concurs with the mean and subject pairing findings. At grade C, the biology line lies above that of physics, which in turn lies above the chemistry line. Thus the relative positions of the lines change for biology and physics across the grade range A*-C making it difficult to suggest which of these subjects can be interpreted as less or more severely graded than the other. Figure 4.2 for the 1995 (Tier 02) population only has readings for all subjects across the B-E grade range within which the associated lines cross each other in several positions. Consequently, no relationships in severity of grading are shown.

The graphical treatments provide a relatively easily assimilated picture of the different science subjects' apparent severity of grading. My findings suggest that based on the assumptions that the analyses inform about 'gradeness ', the science subject shown to be the most severely graded varies from population to population and follows the trend identified from the means in my earlier findings. However, even if one agrees with examining groups that this method informs about 'gradeness', it is shown by the 1995 data to be unhelpful in situations when data across the whole grade range being considered is incomplete and/or when the graph lines intersect in several places. Another reservation I have about this method for investigating examination comparability is that the students in each of my populations will not have taken the same subjects at GCSE and therefore their average grade values will be constituted differently. Taking the average of the average grade values supposedly overcomes this issue – but my view is that differences in the GCSE subjects taken by students must reduce the validity of any such deductions. This use of students' average GCSE grade scores as a baseline for the comparison of performances in individual GCSE subjects is similar to the use of a reference test in other comparability studies (Nuttall *et al.*, 1974) and I have raised several caveats about that in Chapter 3.

4.5 Exploring relationships between students' WJEC science performances, average GCSE grade scores and English and mathematics GCSE performances

4.5.1 Correlation

Spearman correlation coefficient values and their significance were calculated for each population from the primary datasets. Some of the students in these primary dataset populations had missing English and mathematics grades. This is the reason for the relatively low N values in the correlation coefficient calculations, for example N equals 510 for the mathematics and 588 for the English pairings compared with the study's 1993 primary dataset's population of 631 for biology, chemistry and physics. All N values are given in Table 4.5 together with coefficient and significance values.

Respectively for the 1993 and 1994 primary dataset's populations all, and all but five and four of the 'missing' English and mathematics achieved GCSE grades described above were later traced. Tracing the unrecorded grades of the 1995(03) and (02) primary dataset populations proved more problematic. None of the students with missing grades in the 1995(03) population could be traced and seven English and four mathematics missing grades also remained untraceable for the 1995(02) population[1]. A correlation study similar to that listed in Table 4.5 was carried out on the English and mathematics grades for the 1993, 1994 and 1995 (02) datasets that contained all traceable English and mathematics grades. Due to the relatively large number of untraceable missing English and mathematics grades, this process was not repeated for the 1995(03) dataset. There is very little difference in the Spearman correlation coefficient values for each primary dataset's population and its counterpart containing all traceable English and mathematics grades. Therefore any trends identified from the Table 4.5 correlation coefficient values still apply.

[1] WJEC would not allow me to establish the reason(s) for the missing English and mathematics grades data against any of the possible causes identified in Chapter 3.6.4. For example they would not give me access to other examining groups' datasets.

Table 4.5 Correlation Coefficients between the Students' English and Mathematics Grades and their Biology, Chemistry, Physics and Average GCSE Grades - WJEC.

	Biology	Chemistry	Physics	Average GCSE Grade	Mathematics
1993					
English	0.32	0.34	0.35	0.74	0.38
Mathematics	0.37	0.55	0.65	0.69	
1994					
English	0.35	0.37	0.33	0.71	0.30
Mathematics	0.46	0.52	0.58	0.63	
1995(03)					
English	0.42	0.40	0.26	0.75	0.29
Mathematics	0.34	0.51	0.46	0.55	
1995(02)					
English	0.33	0.27	0.26	0.75	0.35
Mathematics	0.46	0.44	0.52	0.64	

All values are significant to the 0.1% level.
(i) For the English and mathematics pairing N = 510 (1993); 697 (1994); 344 (1995/03); 537 (1995/02)
(ii) For pairings involving the variable mathematics N = 510 (1993); 697 (1994); 344 (1995/03); 549 (1995/02)
(iii) For pairings other than (i) involving the variable English N = 588 (1993); 787 (1994); 382 (1995/03); 594 (1995/02).

The following discussions refer to the WJEC primary dataset populations' examination data only. The rationale for this decision is the need to form a reference basis in connection with the assessment practices of different examining groups. For all of the populations a positive correlation between the variable pairings occurs significantly (P 0.001) for: English with each of biology, chemistry, physics, average GCSE grade, mathematics; and, mathematics with each of biology, chemistry, physics, average GCSE grade. One interpretation, which is that of examining groups and others upholding the assumptions of the technical approach to comparability, is that if students perform well in their GCSEs and in each of biology, chemistry and physics, they also tend to perform well in *both* their mathematics and English GCSEs. Those supporting the technical approach with its assumption of the existence of ability would expect a more positive correlation between the English and mathematics GCSE grades (r_S values range from 0.38 to 0.29). Alternatively, on the basis of an educational achievement model, I would not expect them to correlate more positively unless they both had high linguistic or high mathematical demands.

Although both English and mathematics correlate positively and significantly (P=0.001) with each of the biology, chemistry and physics GCSE achieved grades, all of the correlations, except that for 1995 (03) biology, are more positive for mathematics than for English. It would seem a reasonable supposition that this may be explained by the science examination papers requiring mathematical rather than literacy competency to a significant degree . Across populations the most positive correlation was between mathematics and physics achieved grades. As a teacher and lecturer talking to fellow teachers, employers and parents during the past forty years, I have found that physics is perceived as the more 'mathematical' science compared with biology and chemistry and biology the least 'mathematical'. Physics and biology are also respectively viewed as being assessed with examination papers containing proportionally the most and least mathematical questions. The positive correlation between the students' achieved mathematics grades and science grades is from biology (least positive) to chemistry, to physics (most positive), which corresponds to these perceptions. The findings from this correlation investigation were identified as being additional considerations for illuminating 'comparability' when I engage with teachers.

4.5.2 Kappa analysis

Kappa values were calculated for the pairing of English and mathematics and for each of these subjects with biology, chemistry and physics. Adjustments to the primary datasets to ensure the paired subjects held the same awarded grades, and thus allowed kappa calculations, were necessary for all populations except 1995(03), which had its students being awarded all the same grades across the paired subjects (see Chapter 3.6.6).. The kappa values are shown in Table 4.6.

Application of the Landis and Koch(1977) scale for interpreting kappa values indicated an overall higher level of agreement between the grades awarded in the science subjects and mathematics (fair) than for English (slight). This pattern is obtained from comparing the overall mean kappa values calculated from my populations. It also applies for all of the individual populations. Examining group personnel might interpret these outcomes as mathematics showing

Table 4.6 Kappa Values for Biology, Chemistry, Physics, English and Mathematics – WJEC

	Biology	Chemistry	Physics	Mathematics	Mean
English					**(BCP)**
WJEC 1993	0.125	0.117	0.124	0.143	0.122
" 1994	0.072	0.104	0.092	0.100	0.089
" 1995 (03)	0.069	0.138	0.065	0.109	0.091
" 1995 (02)	0.101	0.110	0.049	0.082	0.087
Mean	0.092	0.117	0.083	0.109	
Mathematics					
WJEC 1993	0.116	0.333	0.164		0.204
" 1994	0.101	0.154	0.199		0.181
" 1995 (03)	0.115	0.288	0.212		0.205
" 1995 (02)	0.161	0.189	0.198		0.183
Mean	0.123	0.241	0.193		

Mean (BCP) refers to a particular population's biology, chemistry and physics mean kappa value.
All kappa values are significant to the level 0.1%.

more similar severity of grading with the science subjects than English. My interpretation is that this is not necessarily so and that the sciences and mathematics are requiring students to interact with these assessments in more similar ways than the science subjects and English. The overall mean kappa values show the highest level of agreement to exist between mathematics and chemistry, closely followed by mathematics and physics. However, this pattern is not sustained for each of the study's individual populations. The agreement between mathematics and biology is the lowest of the mathematics/science relationships and this is consistently shown for each of the study's populations. Similarly, the agreement between English and physics is the lowest of the English/science relationships, although this overall trend is only replicated within the 1995(03) and (02) populations. The agreement between the English and mathematics grades is seen to lie consistently within Landis and Koch's (1977) 'slight agreement' range for all populations under scrutiny.

The kappa outcomes show that in terms of being awarded the same grade there is more chance of students doing equally well in mathematics and the science subjects rather than English.

This supports an educational achievement model of the sciences as being predisposed towards mathematics than English. Mathematics and biology consistently show the lowest level of awarded grade agreement, which could be argued supports the perception that biology is the least mathematical of the sciences as described in the correlation studies above. Even so, my analyses are only for these selected populations which are sampled from whole GCSE subject student populations and are not representative of these. The outcomes therefore serve to illuminate *potential* interrelationships rather than establish causal relationships.

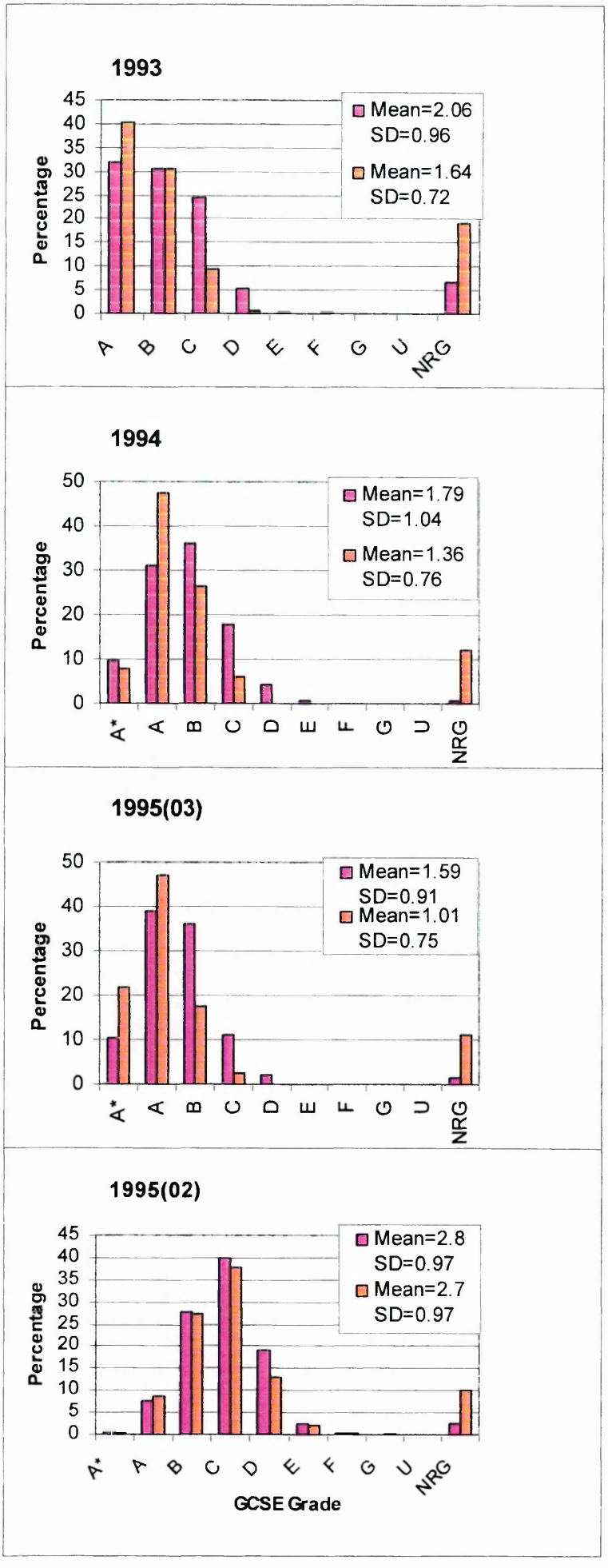
4.5.3 Descriptive Statistics

A greater understanding of the different populations' relative attainment in both English and mathematics was sought to explore the correlation and kappa findings. The percentage distributions across grades for the different populations' English and mathematics achievements and the means and standard deviation values are given in Figure 4.3.

Two main trends are apparent. First, the mean values for English tend to be higher than the population's corresponding mathematics' mean value. Given the reverse nature of the grading scale, this indicates that each of the populations of this study achieve overall better grades in mathematics than in English. The technical approach to comparability would argue that this shows English as apparently being more severely graded than mathematics. An alternative interpretation is that my populations were better at meeting the skill demands of the mathematics rather than the English assessments. Second, the dispersion of achieved grades is greater for English than mathematics for all populations except that of 1995(02) where the standard deviation values are the same. Figure 4.3 shows that for each of the populations, there is a tendency for the students' achieved mathematics grades to be more positively skewed than their achieved English grades and with a tendency for students to achieve proportionally more of the examinations' two top grades in mathematics than in English.

This pattern holds for all four populations. It is tempting, given the prior findings, to conclude

Figure 4.3 English and Mathematics Grade Distributions - WJEC



that students who perform well in mathematics GCSE have the skills to respond well to the mathematical requirements of the science GCSE subjects.

The findings discussed have highlighted the potential for the requirements of assessment items, for example mathematical-based tasks, to influence students' performances in different subjects, an issue that is largely ignored in the technical approach to comparability. It seems timely to turn now to my previous research on the cognitive skills of the study's examination papers and view these findings against those presented thus far in this Chapter.

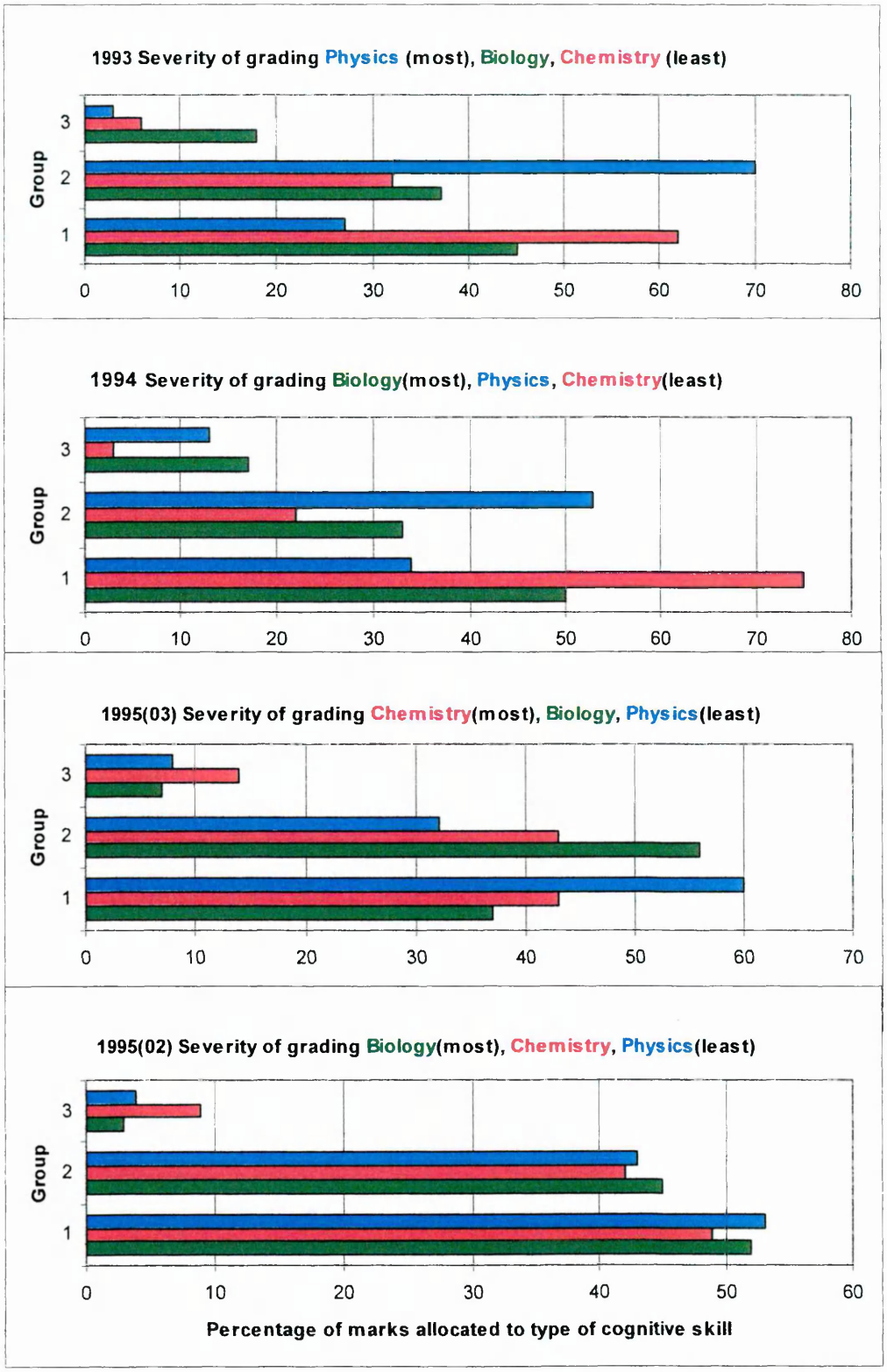
4.6 Is there any quantitative relationship between performance and cognitive skill demands of examination papers?

Appendix 2 shows the findings from previous research (Benson, 1995) on the cognitive demands of the WJEC examination papers associated with the current study. Figure 4.4 illustrates these mark weightings for the different groups of cognitive demands *and* the science subjects' order of grading severity identified from the subject-pair method.

From Figure 4.4 for all of the populations in this study, the science subject in which students achieved the relatively higher grades i.e. with the least apparent severity of grading, had the majority of its examination papers' marks being weighted to recall of knowledge. For example, chemistry is the least severely graded science subject in 1993 and 1994 and in both years also has papers with the highest weighting for recall of knowledge. In particular, this association is seen to hold as physics markedly changes its order of severity of grading for the 1995 examinations and in this respect it is tempting to view the technical approach used as valid. However, the order of severity of grading does not appear to be associated with a particular order of weighting for any of the other cognitive demands for all three subjects and there is no apparent relationship to associate with the high positive correlation of the physics and chemistry performances (see 4.3.2). The 1995 examinations in both tiers have a more even distribution of mark weightings across the different cognitive demands than those of 1994 and 1993. This coincides with the first cohort of students having followed the National Curriculum being examined in 1995 with new WJEC syllabuses and it is tempting to suggest that my findings are reflecting changes in the nature of the items related to assessment artefacts. If this is so, a

Figure 4.4 Subject Grading Severity (subject pair method) and Examination Paper Cognitive Demand

Group 1 = Knowledge
Group 2 = Comprehension and Application
Group 3 = Analysis, Synthesis and Evaluation



technical approach that merely compares examination grade distributions across time is highlighted as not taking such events into account.

In terms of comparability, arguably Figure 4.4 highlights a strong association between the examination papers' recall of knowledge weighting and severity of grading and this is how examining group personnel would interpret the analyses. However, the analysis presented in Figure 4.4 isolates one variable within a multiplicity. For example, no account was taken of the contexts within which the items (questions) were based, nor of the change in coursework for all science subjects in 1995, the differences in the science subjects' coursework arrangements prior to this and the potential these coursework issues might have on students' science grades. The analyses therefore merely points to a potential influence. Furthermore, the allocation of examination questions to particular types of cognitive skill was the result of PGCE students' subjective judgments being used to provide a consensus view for the final allocation of each question. Importantly, as argued earlier, if a constructivist model of learning and assessment is adopted, the demand emerges in the interaction between students and the questions – what cognitive skill *they* used to answer any particular question. The cognitive demand weightings in Figure 4.4 are only illuminative. Once again, the complexity of exploring examination comparability is highlighted along with the shortcomings of the traditional technical approach. The change in the examinations' cognitive demand weightings across time may emerge as an issue when I engage with teachers and may provide further insights on 'gradeness'.

4.7 Are there relationships between sex group and achieved WJEC science, English, mathematics performances and average GCSE grade scores?

4.7.1 Inferential and descriptive statistical analysis.

As stated in Chapter 3, in this section I consider biological sex groups because that is how the data is presented. However, in interpreting the findings I am aware that they represent overall rather than individual effects. All of this study's populations contain a substantial majority of boys (Table 4.7) in line with entry patterns for Triple Award GCSE Science (Murphy and Whitelegg, p. 41, 2006).

Table 4.7 Number and Percentage of Boys and Girls in the Study's WJEC Populations.

	Population							
	1993		1994		1995(03)		1995(02)	
	n	%	n	%	n	%	n	%
Boys	368	62.1	446	64.3	231	59.7	362	59.5
Girls	220	37.9	251	35.7	156	40.3	247	40.5

Willingham and Cole (p. 98, 1997,) note that the relative number of boys and girls taking an examination can be an important consideration in describing and understanding 'gender' difference and similarity in selected groups but that *'the nature of the selected group depends partly on the character of the selection and that ordinarily there is little precise information about that process'*. Here, it is important to note that the sex sub-groups in my populations are in a sense *'restricted samples'* (ibid.). My populations consist of boys and girls who have chosen to study Triple Award Science GCSE. That choice will have been influenced by many factors, for example the personal wishes of the students, the advice of their teachers, and parental pressure. All of the boys and girls in my populations have been entered for the same tier of examinations but entry decisions will again have been influenced by numerous factors. There is little precise information about the selection processes for students taking Triple Award Science GCSE courses and then their GCSE tier entry, in accord with Willingham and Cole's quotation above. These 'selection' processes and their consequences for comparability are an intended focus when I engage with teachers in the qualitative dimension of my research. In this Chapter my intention is first to investigate the boys' and girls' performances, acknowledging that I do not know what might be the effects of sample restriction on those performances. Statistically, as Willingham and Cole note and found, if there is a selected group (restricted sample) then there would be an expected difference in the standard mean difference between boys and girls in favour of the minority sample. In each of my populations girls form the 'minority' group, with between 38 and 40 per cent of each WJEC population in my study being composed of girls. Willingham and Cole would expect the girl's sub-

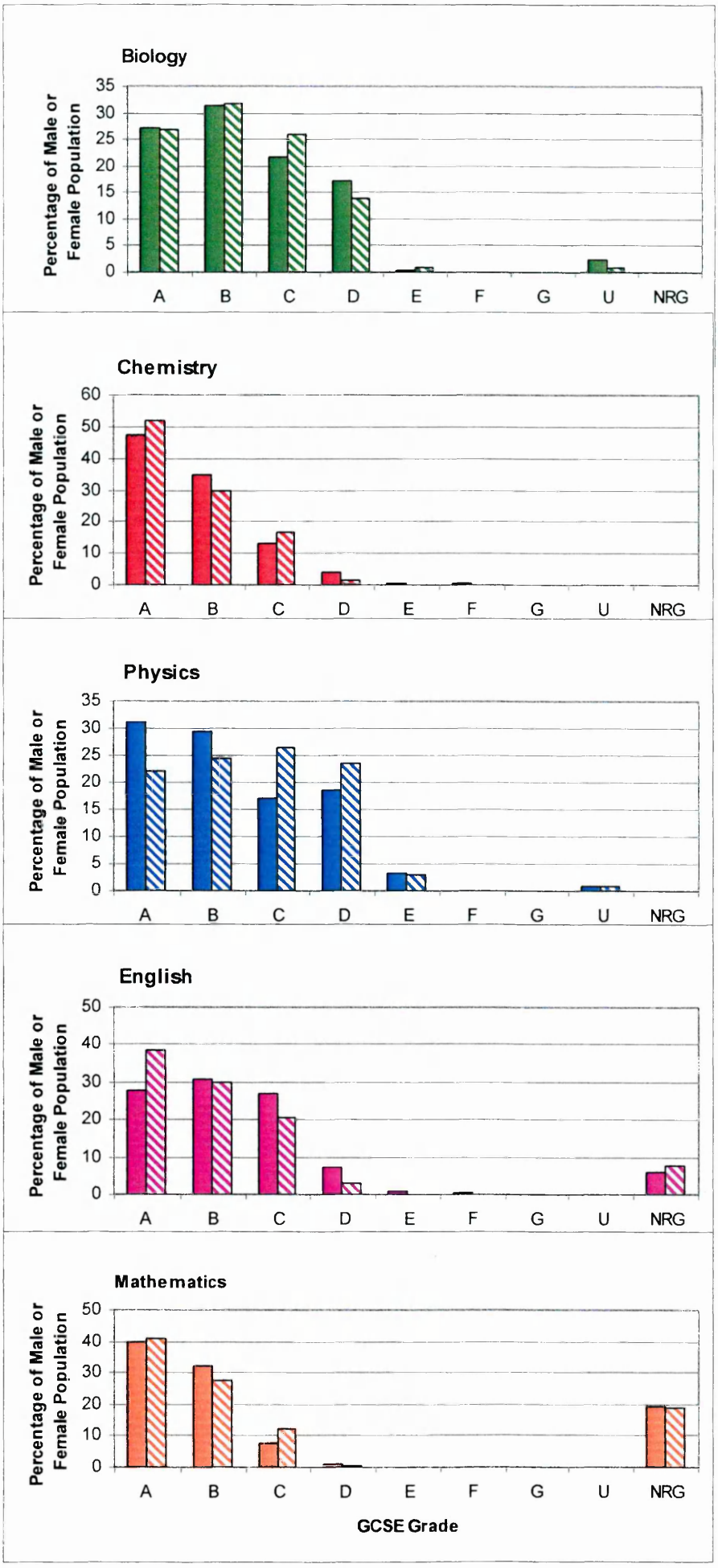
group in my populations to perform better than that of the boys. I take their proposition into account when interpreting my analyses.

The only other published study of girls' and boys' performances at GCSE in Wales focuses on 1992 to 1997 data and so includes the years of my quantitative data (1993 – 1995 inclusive). This study was commissioned by the Qualifications, Curriculum and Assessment Authority for Wales (ACCAC) and published in 1999. It draws together data from across the whole of Wales covering attainment in national assessments from age 7 to 19. The statistical analysis is based on statutory assessment and examination results, namely Key Stages 1, 2 and 3, GCSE and 'A' level. It focuses on the 'gap' in achievement between boys and girls at school for each subject and for each attainment and grade level. The study defines an achievement gap as the difference between the performances of boys and girls, taking account of the patterns in entry. This approach is adopted for fulfilling the study's primary focus which is on *changes* in entry and achievement across time. Unlike my research the ACCAC study does not take any account of tier entry in its examination populations and their GCSE results and bases its calculations on whole examination entries for specific GCSE subjects. Only achievement gap calculations are used and so, for example, there are no correlation, kappa or inferential statistical analyses to compare with those of my research. The study also only presents its analyses on sex differences for science as a whole, claiming that there is a small achievement gap in favour of boys at the higher grades and that *'gender is not a clear problem in science'* (p. 30, 1999). These claims subsume any differences in sex achievements for specific awards, for example Triple Award and Double Award and little is said about sex differences in the different science disciplines within each type of award. There are general statements that there is an achievement gap in favour of girls in biology and no achievement gap in chemistry and physics but no statistical analyses results are presented to substantiate these statements or indicate the extent of the gap. My study attempts to control for tier entry and deals only with biology, chemistry and physics in GCSE Triple Award. Comparing my results with those of this ACCAC publication are therefore of limited use but reference is made where relevant.

My sub-groups' achieved GCSE grades were first investigated with inferential statistics. The results were then explored with descriptive statistics. Cross-tabulations of sex with each of the variables biology, chemistry, physics, English and mathematics grades in the form of frequencies for each grade category are illustrated in bar charts (Figure 4.5a-d). This approach was repeated with each of my populations. This analysis was based on percentages of each sub-group's population rather than the population as a whole (see Chapter 3) to increase the validity of my comparisons. This method is seen by Gorard, Salisbury and Rees (p. 7, 1999) in their ACCAC study as appropriate for identifying the presence or absence of patterns in differential achievement, although they advocate their 'achievement gap' calculations for determining differential sex achievements across time. My main concern is not with comparing achievements across the three years of my study to identify precise relative percentage changes but with patterns in sex differences within each year of my data and their statistical significance. As discussed in Chapter 3 there are many influences on examination performance including many which emanate from the nature of the examination population itself. For this reason alone I question the validity of using precise percentage 'achievement gap' changes for reporting on boys' and girls' GCSE performance changes across time. Thus I looked for similarities and differences across the years but not for specific percentage changes as required by the achievement gap calculations (Gorard *et al.*, 1999). Furthermore, unlike the achievement gap method, my chosen calculation method is used widely for reporting on sex, social class and ethnic group differences in achievement (Robinson and Oppenheim, 1998; Bright, 1998; Bentley, 1998; Gillborn and Gipps, 1996; Murphy and Whitelegg, 2006) and so allows comparison of my results with those of other researchers.

Using the language of examining groups, my results indicate there were no significant differences between the 1993 boys' and girls' achieved biology, chemistry and mathematics grades. However, the boys performed significantly ($P = 0.008$) better than the girls in physics and proportionally more boys than girls achieved grades A and B (see Figure 4.5). In chemistry, the degree of positive skewness is more marked for both boys and girls than in the other two sciences, normal distribution being lost. This concurs with chemistry being overall the least severely graded or

Figure 4.5a Distribution of Boys' and Girls' Grades – 1993, WJEC
Block of colour = Boys; Hatched colour = Girls

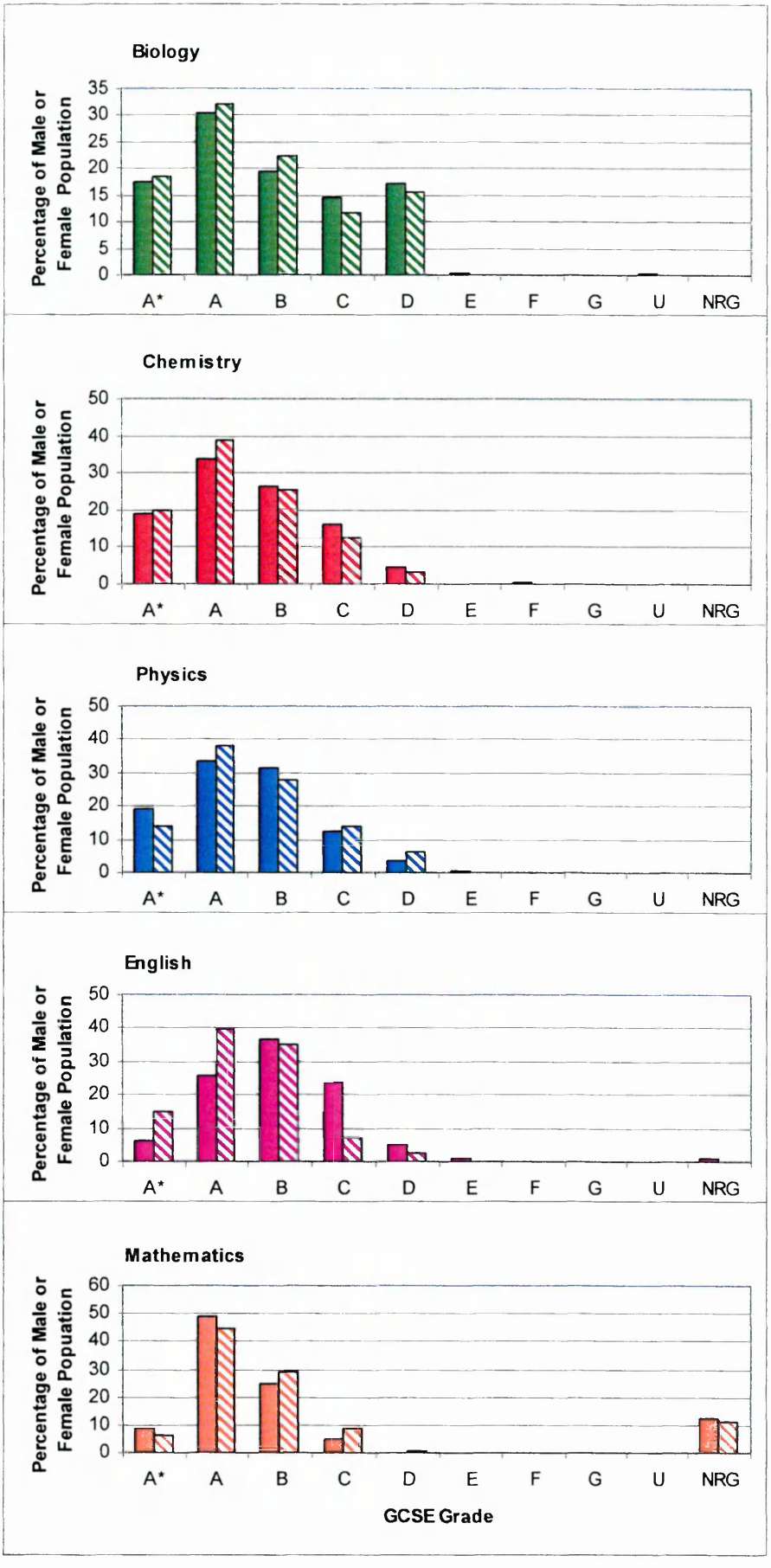


the science subject where students overall achieved relatively more highly in the three science subjects for the 1993 population. Girls performed significantly ($P = 0.001$) better than the boys in English and in their average GCSE grade scores. The greatest disparity in the proportion of achieved English grades occurs for grade A, with 38.5 per cent of the girls' sub-group achieving grade A compared with 27.8 per cent of the boys' sub-group.

Figure 4.5 shows that the distributions of mathematics grades for the boys' and girls' sub-groups are similarly positively skewed, with a slight tendency for the boys to proportionally outperform the girls at grade B and this situation being reversed for grades A and C. There is no statistically significant difference between the girls' and boys' mathematics grades. From these results, it would appear that there is no sex effect between achieved GCSE mathematics and physics grades. Separate correlation studies of the mathematics and physics grades achieved by the boys and the girls in the 1993 population were carried out to explore further the significant differences between the sub-groups' physics grades. Both the boys' (0.68) and the girls' (0.62) Spearman correlation coefficient values for mathematics and physics grades are significantly (P=0.001) positive. One would therefore expect for both sexes high achievement in mathematics to be associated with high achievement in physics. The correlation coefficient values are sufficiently similar as not to offer any insights as to why the girls achieve significantly less well than the boys in physics. I interpret these technical findings as indicating population / subject interaction, in that something(s) within the physics examination and/or coursework evoked a reaction from individual students that taken collectively manifests itself as a sex sub-group effect. This interpretation challenges the notion of any of the sciences being described as more severely graded with the question 'more severely graded for whom?'.

The inferential statistics suggest there are no significant differences in the 1994 boys' and girls' achieved physics grades. However, the boys achieved significantly ($P = 0.008$) better grades in mathematics than the girls. In turn, the girls achieved significantly ($P = 0.001$) better English grades and average GCSE grade scores, the latter showing an even greater difference than in 1993 (mean difference is 0.1 in 1993 and 0.27 in 1994). Technically this average GCSE grade score suggests the

Figure 4.5b Distribution of Boys' and Girls' Grades – 1994, WJEC
Block of colour = Boys; Hatched colour = Girls



1994 girl's sub-group were more able at meeting the demands of their GCSE subjects than their 1993 counterpart with the technical approach's assumption that grades will have common currency between subjects and across years. For me this is not necessarily so because of a multitude of possible influences. For example: the 1993 and 1994 examinations may have required differing skills and there could be differences in the girls sub-groups' interactions with these; different subjects with their specific skill demands may have been taken by the two sub-groups so that the same subjects are not being compared; marking *may* have been more lenient in one or more subjects in 1994, despite examining group's arrangements that claim to stabilize this. Continuing with an examining group's type of interpretation and within my chosen value of 5 per cent for significance, girls achieved significantly ($P = 0.048$) better in chemistry than the boys. There was no significant difference in achievement between the boys and girls in biology. Biology has been identified in the subject pair method as the most severely graded science GCSE subject for the 1994 population of this study. It therefore appears that for both boys' and girls' sub-groups this was the case.

The significant differences in the 1994 boys and girls sub-groups' achievements in chemistry, English and mathematics were explored further by conducting correlation studies. Both the boys (0.54) and girls (0.52) sub-groups' Spearman correlation coefficients for mathematics and chemistry grades are significantly (P=0.001) positive. Therefore, for both groups, high achievement in mathematics is associated with high achievement in chemistry. The correlation coefficient values for the two groups are sufficiently similar to indicate that girls' significant better performance in chemistry is not related to the relative performances in mathematics GCSE, which suggests to me that the underlying issues are more complex than just the mathematical skill demands of the examinations. The Spearman correlation coefficients were recalculated for the 1994 boys (0.52) and girls (0.48) sub-groups using the dataset that included all available mathematics achieved grades (506 of the total 509 males; 283 of the total 283 females), regardless of examining session and examining group. The correlation coefficients differ by a similar relatively small amount (0.04) to that obtained from the incomplete mathematics dataset (0.02) providing no further insights to explain the differential sub-groups' achievements in chemistry in terms of awarded GCSE mathematics' grades. As noted earlier in the discussion about physics, the

tier of mathematics examination paper for which the boys and girls are entered might be one contributing factor.

In discussing the relationships between the 1994 boys and girls sub-groups' chemistry achievements, it again seemed useful to explore the interplay between the boys' and girls' chemistry and English achieved grades. Consequently, Spearman correlation coefficients for the separate sub- groups' English and chemistry grades were calculated. The English grades achieved by girls correlate less positively (0.24) with their achieved chemistry grades than is the case for the boys (0.43). One way of interpreting the correlation coefficients is to say that for the 1994 population, high achievement in chemistry is more likely to be associated with high achievement in English for the boys' than the girls' sub-group. Nor is it valid, on the basis of the evidence available at this point, to claim that the girls' sub-group achieve well in chemistry because they are high achievers in English GCSE. Nevertheless, despite the girls sub-group's relatively higher achievements in both English and chemistry compared with those of the boys' sub-group, the positive association of achievements in these two subjects appears to be less marked for the girls than the boys.

For the 1995(03) population there were no significant differences between the boys' and girls' achieved biology, chemistry and mathematics grades. Overall, boys achieved significantly ($P = 0.007$) better grades in physics than the girls. Conversely, the girls achieved a significantly ($P = 0.000$) better average GCSE grade score and significantly ($P = 0.018$) better English grades than the boys. There are no significant differences in the boys' and girls' achieved biology and mathematics grades for the 1995(02) population. Boys out-performed girls in physics but not to such a significant level ($P = 0.08$) as their 1995(03) counterparts. Overall, the 1995(02) girls' sub-group achieved significantly ($P = 0.042$) better than the boys' sub-group in chemistry and even more significantly ($P = 0.000$) in their average GCSE grade scores. The girls' sub-group also performed better than the boys in English with the significance becoming greater from 1995(03) ($P = 0.018$) to 1995(02) ($P = 0.000$).

Correlation studies were conducted to explore further the statistically significant differences in the boys and girls sub-groups' achievements in physics and English for Tier 03 and in physics, chemistry and English for Tier 02 of the 1995 examination session. The correlation coefficients for

Figure 4.5c Distribution of Boys' and Girls' Grades – 1995(03), WJEC
Block of colour = Boys; Hatched colour = Girls

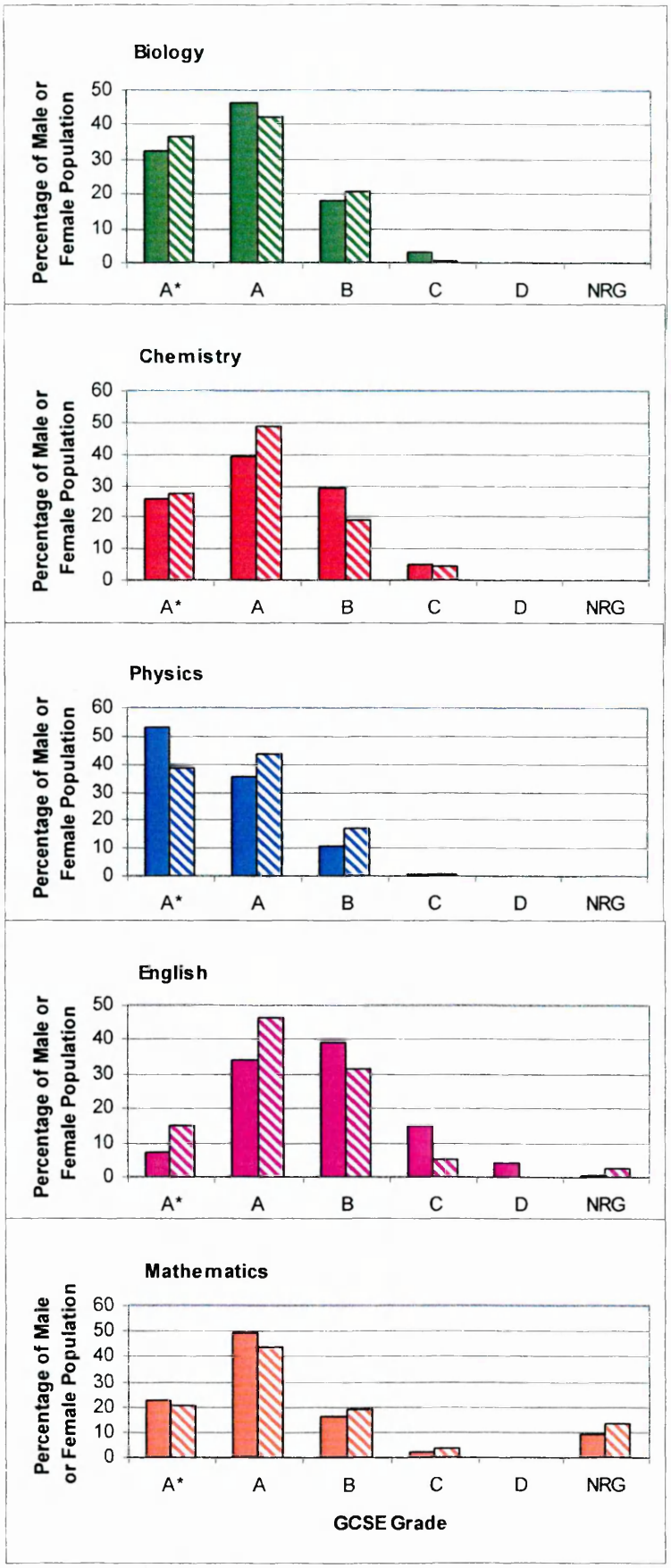
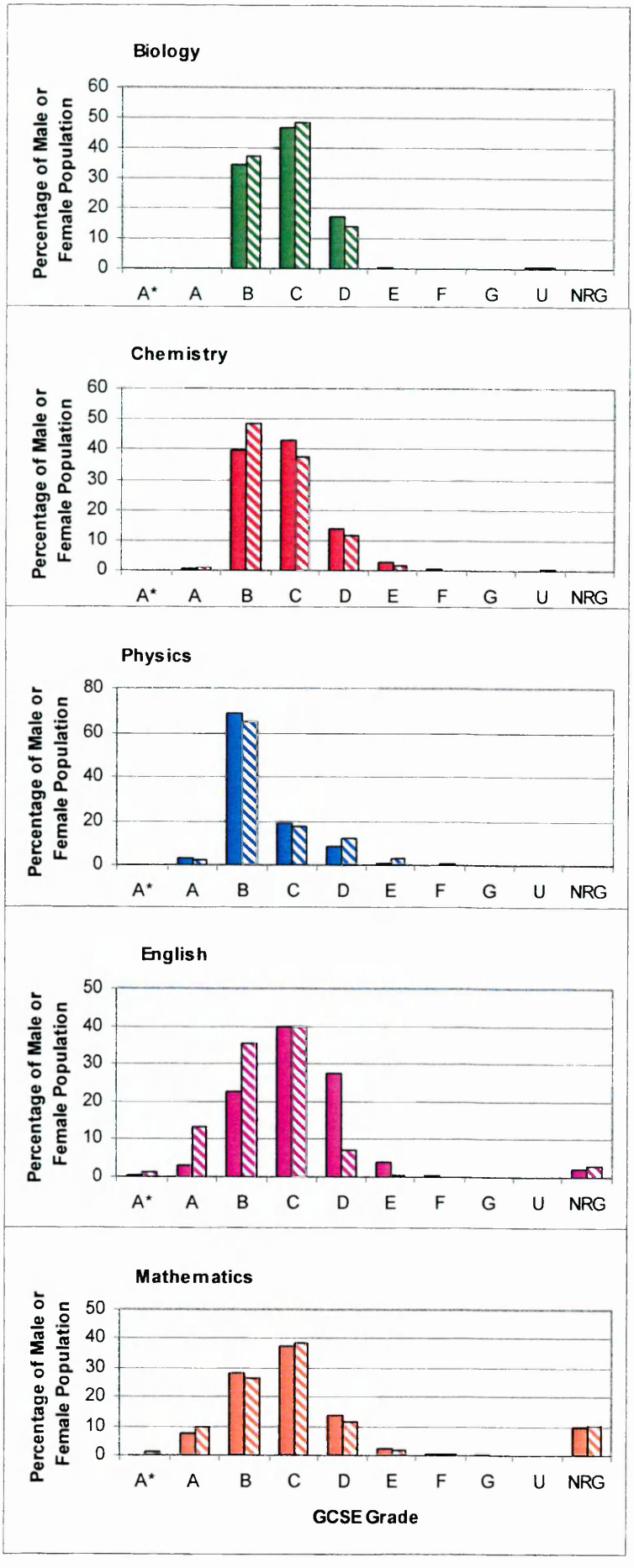


Figure 4.5d Distribution of Boys' and Girls' Grades – 1995(02), WJEC
Block of colour = Boys; Hatched colour = Girls



English and physics (boys = 0.24; girls = 0.36) achieved grades are significantly ($P = 0.001$) positive. Therefore, high achievement in physics is associated with high achievement in English for both sub-populations. However, the correlation values lie in the 'weak' correlation range (Coolican, 1994), with that for the girls being only slightly more positive than that for the boys. Using the language of examining groups, there is only a slight tendency for the girls' high achievement in English to be more predictive of high achievement in physics than is the case for the boys. The values for both sub-groups are sufficiently similar so as not to offer an explanation as to why the male sub-group achieves significantly better physics grades than their female counterpart. The same deductions apply for the physics: English correlation coefficient values for the 1995(02) male (0.28) and female (0.36) sub-groups. Furthermore, the 1995(02) correlation values for the English : chemistry pairings are sufficiently similar (boys 0.28, girls 0.23) in their degree of positive correlation as to not provide insights into the differential performance of the two 1995(02) sub-groups. For the chemistry: physics pairings the values (boys 0.53, girls 0.60) indicate that both 1995(02) male and female sub-groups tend to have their high achievement in chemistry moderately (Coolican, 1994) positively associated with high achievement in physics, where this is slightly more so for girls than boys.

The ACCAC study showed that *'girls gain more of the higher attainment levels at Key Stages 1 to 4 in English'* (p. 5, Gorard *et al.*, 1999) but in the study's 1999 publication no reference is made to significance levels. My findings also show girls significantly out-performing boys in English and gaining relatively more of the higher grades in all three years of my data. ACCAC also showed girls out-performing boys in biology. However, my results consistently show no statistically significant difference in the boys' and girls' performances in biology. Unlike ACCAC, I have found that boys significantly out-perform girls in physics. This applies for three out of four of my datasets and even in the dataset where this does not apply, boys gain more of the highest grade (A*) than girls. This finding is shown to hold in more recent studies. Data relating to examinations from 2000 to 2004 (Murphy and Whitelegg, 2006) shows that in England far fewer girls than boys are entered for physics in Triple Award and their performance relative to boys is lower across the pass grades. Indeed, overall the 2000 and 2001 GCSE results showed that physics (at Triple Award) was the only subject where boys

achieved a higher proportion of A*-C, the pass grades, than girls. Across 2000-2004 boys also gained proportionally more A* and A grades than girls in physics and mathematics. However, more boys than girls are entered for the higher tier paper in mathematics, which allows access to these two grades (Elwood and Murphy, 2002). Because of the restricted and selected sample it suggests that girls are not achieving as well in physics in relation to boys. This 2000 – 2004 performance trend is also shown in Wales where there is a similar though proportionally smaller discrepancy in entry between boys and girls and yet boys out-perform girls marginally (ibid). Scotland and Eire have different 16+ assessment systems to those in England and Wales. Nevertheless, entry to 16+ physics examinations in Scotland and Eire shows a similar gap in favour of boys to that in England and Wales. However, girls outperform boys on the top grades, which could be expected for these restricted samples in contrast to girls in England and Wales (ibid.). Here it could be assumed that each smaller sample of girls is more highly selected and therefore more able and motivated persons predominate amongst the girls than the boys (Willingham and Cole, 1997).

Preece *et al. 's* (1999) analysis of the 1996 Key Stage 3 science test results for a large representative sample of schools found that the largest gaps in performance occurred on questions assessing physics, where boys outperformed girls. When this study was repeated with Key Stage 3 2003 results with t-tests on the mean scores achieved by boys and girls, boys were again shown to significantly outperform girls on physics on both the lower and higher national science tests. The publication of Key Stage 3 test results by level and overall points achieved across papers and subject components masks these findings. In this respect the same reservation I have discussed earlier about sub-group effects being masked in an examination population's GCSE grade distributions is raised. The international survey of students at grade 8 (approximately 14 years), Trends in International Mathematics and Science Study (TIMSS), show boys out-performing girls on the physics curriculum content of science performance measures (Ruddock *et al.*, 2003), but with the gap between girls' and boys' performances decreasing from 1995 to 2003. However, it should be noted that the Key Stage 3 national tests in England with their emphasis on structured responses are different in item style and content to those used in TIMSS surveys. The Programme for International Student Assessment (PISA)

assesses the scientific literacy of 15-year olds and so assesses students at an age closer to that for students taking GCSE than the Key Stage 3 tests and TIMSS surveys. The test items focus on the application of scientific knowledge and skills to real-life situations. The PISA results reveal no gender differences but then they are based on tests which assess different skills and understanding and use different item formats to the TIMSS surveys; multiple choice items dominate TIMSS assessment whereas PISA uses general item formats including structured and open written responses.

Returning to the ACCAC study (1999), it found no achievement gap in mathematics. However, my 1994 and 1995(02) datasets show boys significantly out-performing girls. Similarly, ACCAC found no clear achievement gap in chemistry but my 1994 and 1995(02) datasets show girls significantly out-performing boys. ACCAC (1999) does not provide numerical results or indeed details of the sizes of examination populations they have used in their calculations for any of the science subjects. Entries for all of the GCSE science subjects are summed to give an *'overall picture'* (p. 52, 1999) and there are just statements about girls' and boys' relative achievements in biology, chemistry and physics. For this reason and that I have attempted to control for tier entry by only considering candidates entered for the same tier of biology, chemistry and physics examinations, I regard my study as offering additional insights into boys and girls relative performances in Wales than that of ACCAC. Differences in my findings with those of ACCAC could result from the differences in our chosen methods. How 'comparability' is defined and consequently what is used to measure it are again highlighted as contentious key issues.

4.7.2 Sex sub-group comparability

Using the language of examining group personnel, the technical investigations suggest that the female sub-groups of all of the study's populations consistently out-perform, and at a statistically significant level of 0.1%, the associated male sub-groups in their GCSE subjects as measured by their average GCSE grade scores. This finding concurs with reports of girls performing better than boys at GCSE in the national press across recent years (Independent, 5.1.98; TES: 17.6.05) and figures released by the Joint Council for Qualifications on their web site, where these reports are based on figures presented as percentages of boys and girls obtaining: five or more subjects with grade A* to C; five or more

subjects with grade A* to G; and from 2006, five or more subjects with grade A* to C including C or better in both English and mathematics. My findings also show that the female sub-groups have *increasingly* out-performed their corresponding male sub-groups across the examination sessions of the study. When the means of the average GCSE grade scores from both sub-groups are compared (girls' mean – boys' mean), they differ by the equivalent of: nearly a quarter (0.22) of a GCSE grade in 1993; over a quarter (0.27) of a grade in 1994; nearly a third (0.30) of a grade in 1995(03) and nearly half of a grade (0.46) for the 1995(02) population.

However, the caveats I raised about the technical approach to comparability in Chapter 3 still apply. For example I could argue that the girls of the various populations have sat a different combination of subjects than the corresponding boys and this could invalidate comparisons of each sub-groups' average GCSE scores. The sub-groups have been treated homogeneously because the data was presented to me in that form – no account was taken of differences in ethnic origin or socio-economic factors within each sub-group or across the three years of the investigation. As discussed in Chapter 3, these factors have been shown to influence examination performance (Smith and Tomlinson, 1989; Nuttall *et al.*, 1989; Drew and Gray, 1990, 1991; Troyna, 1991). Changes in the structure of examinations can also change a pattern in girls' and boys' assessment performances (Gipps and Murphy, 1994; Elwood, 2001). The structure has changed because coursework arrangements, syllabus and available grade ranges have all changed during the three years of my study – years that also fall within the ACCAC (1999) study and rendering it also open to criticism because of the caveats I have raised.

Continuing with a technical interpretation, my findings include all of the female sub-groups consistently out-performing their male counterparts in GCSE English. This occurs at a significance level of respectively 0.1% for three and 2% for one of the four populations and well within my chosen significance level of 5%. Overall there is a slight tendency for more girls than boys to have no recorded grade in English. This might be due to more girls than boys being entered for English a year earlier than the rest of their GCSE subjects. One might also speculate that the better female performance in English might be attributed to examination centre tier entry factors, for example a

tendency for boys to be entered for tiers of English papers that do not give access to the higher grades. This is possible as the research of Gillborn and Youdell (1998) and Stobart *et al.* (1992) has highlighted teachers' perceptions of students' abilities and tier entry decisions as determining students' overall achievement. Boys only significantly out-perform girls in mathematics for the 1994 population. In general more girls than boys are entered for the intermediate tier of mathematics (Elwood, 2001). Elwood found that teachers place girls in the intermediate tier to protect them from their lack of confidence and anxiety of being unclassified if their performance drops below the key grade C in the higher tier. Boaler (1997) has shown that the underachievement of 'bright' girls within the higher tier mathematics may be due to the context of the environment in top set mathematics classes, where common features are speed, pressure, competition and reward for getting the correct answers rather than for understanding. As the tier of entry is unkown my findings cannot substantiate these., I am not comparing like with like which also potentially undermines the comparison of students' mathematics – and English, grades. My findings might be reflecting proportionally more girls than boys in my sub-groups being entered for mathematics tiers with no access to the top grades and proportionally more girls than boys being entered for higher tier English. My view is that it is more complex than this and that my findings are more likely to be reflecting several influences, such is the complex nature of comparability.

I have used each science subject's mean value from section 4.3.1 to place them in an order of apparent severity of grading, for each of my investigation's four populations. I have then used the science subjects' mean values for each of my populations' sex sub-groups to place the subjects in an order of severity of grading for each sub-group. Table 4.8 summarizes this information which seeks to reveal any apparent disparities between population and sex sub-group severity of grading.

Table 4.8 Differential Gender Performances, WJEC

Population	1993			1994			1995(03)			1995(02)		
Science Subject	**M+F**	**M**	**F**	**M+F**	**M**	**F**	**M+F**	**M**	**F**	**M+F**	**M**	**F**
Most severely graded	P	B	P	B	B	B	C	C	C	B	B / C *	P
Least severely graded	C	C	C	C	P	C	P	P	P	P	P	B

B = Biology; **C** = Chemistry; **P** = Physics M+F=boys and girls M=boys sub-group F= girls su-group
B / C * denotes Biology and Chemistry as being the same in terms of severity of grading

The trends in science subject grading identified for the different populations are shown to subsume apparent sex sub-group differences. For example, physics was the most severely graded of the sciences for the 1993 population. This is only true for the girls' sub-group, for boys biology was the most severely graded of the sciences. Overall, for the boys' sub-groups there is a greater tendency for biology to be the most severely graded of the science subjects, followed by chemistry. Physics does not appear as the most severely graded of the science subjects for any of the boys' sub-groups of this investigation, even when it is identified as the most severely graded subject for a whole population as in 1993. Conversely, the girls' sub-groups have a tendency to have physics as their most severely graded science subject even when it does not hold this position for the population as a whole, for example 1993 and 1992(02) populations. There is a change in the pattern of science subject severity in the 1995 examinations, at which point chemistry ceases to be the most leniently graded science subject for *both* the boys' and girls' sub-groups. A change in examination performance patterns from 1995 has been highlighted in previous findings, for example the graphical analysis showed a greater similarity in the whole population's performance for the three science subjects for 1995(03) and (02) compared with 1993 and 1994.

This investigation illustrates the limitations of a technical approach to comparability when whole population's performances are considered. Sub-group effects are masked by overall performances. This is shown above for just one type of sub-group, sex. It begs the question what other types of sub-group effects may be masked by just considering a whole population's performance and adds weight to my view of the technical approach and findings being limited in their dependability. Even if one takes the sex sub-group findings as dependable, the issue of treating the sub-groups

homogeneously remains and the counter argument, for example, that not all of the girls found physics harder than chemistry and biology, only some of them did. If there is an issue about physics *per se* and girls underperforming boys, then one would expect to see the sex sub-group differences found in the WJEC populations repeated in the Southern Examining Group (SEG) examination populations of my study. These are now considered.

4.8 How do the findings for the Welsh Joint Education Committee (WJEC) examination populations compare with those of the Southern Examining Group (SEG)?

Data for only two of the three examination sessions relevant to this investigation were made available by SEG, namely for 1994 (for the tier known as 'extended' option) and for 1995 (for the tier known as 'high' option). The SEG 1994 'extended' option was associated with the award of GCSE grades A* - U and the 1995 'high' option with grades A* to C. If students failed to achieve at least a grade C in the 1995 'high' option, the policy was to award them grade U and none of the grades D to G inclusive. The SEG datasets did not contain English or mathematics GCSE achieved grades. Thus the SEG data is only used to investigate relationships between the students' and sex sub-groups' biology, chemistry, physics and average GCSE grade scores and with no reference to either English or mathematics. The findings from the SEG investigations are compared with those of WJEC statistical treatment by treatment to avoid repetition and enhance the flow.

4.8.1 Exploring relationships between students' performances in SEG biology, chemistry and physics?

Subject-pair analysis and findings

Tables 4.9 and 4.10 show respectively the mean grades and the subject pair results. Here again initially I use the expressions of the examining groups in discussing the findings in that differences in grade achievement are first interpreted as differences in severity of grading. However, as I continue to argue, the findings reflect differences in the grades achieved by the populations being compared and how these are interpreted will depend on what is seen to influence students' interactions with assessment items.

Table 4.9 Biology, chemistry and physics means – SEG

	Subject	Mean	
1994 (*N=1001*)	Biology	2.07	Expected grades A* - U apply
	Chemistry	2.08	
	Physics	2.34	
1995 (*N=1759*)	Biology	1.17	Expected grades A* - C apply
	Chemistry	1.36	
	Physics	0.92	

Table 4.10 Subject-pair analysis - SEG

	Mean grade (A)		Mean grade (B)		Difference (A-B)
1994	Biology	2.07	Chemistry	2.08	- 0.01
	Biology	2.07	Physics	2.34	- 0.27
	Chemistry	2.08	Physics	2.34	- 0.26
1995	Biology	1.17	Chemistry	1.36	- 0.19
	Biology	1.17	Physics	0.92	0.25
	Chemistry	1.36	Physics	0.92	0.44

Negative differences indicate (B) is more severely graded than (A)

For the 1994 population there are differences in severity of grading with physics apparently being the most and biology the least severely graded subject. For the 1995 population chemistry is apparently the most severely graded and physics is the least severely graded subject. These subject-pair values indicate there are no patterns in severity of grading between the 1994 and 1995 populations. However, physics changes its position from being the most severely in 1994 to the least severely graded science subject in 1995. Physics also became the least severely graded subject in 1995 for both of the WJEC tiers investigated. Similarly, chemistry appears to become more severely graded for both WJEC and SEG from the 1994 to 1995 examination sessions in my investigation. GCSE examining groups examined newly introduced Triple Award GCSE syllabuses for the first time in 1995. Thus one interpretation is that my technical analysis shows the change in severity of grading of physics and chemistry for both the WJEC and SEG populations as possibly being associated with changes in assessment occurring with the first examination of new syllabuses by GCSE examining groups in 1995. Perhaps the findings are reflecting changes in the ways that the subjects are defined in the syllabuses from 1995, with consequent changes in the ways by which they are examined. For example, physics could have become less demanding in its mathematical skills, more intellectually challenging

concepts might have been omitted in the new syllabuses. The concerns of chemistry teachers about their subject being redefined by the introduction of geological content in the National Curriculum for chemistry referred to in Chapter 1, may be reflected in its apparent increase in severity of grading from 1995. Arguably, because the severity of grading of any of these subjects is shown to vary across time, apparently there is no inherent difficulty associated with a subject itself. For example, my WJEC and SEG populations do not interact with physics to produce consistently lower GCSE grades than in biology and chemistry – physics is not shown as being the most 'difficult' subject *per se*. One might argue that the findings are due to differences in the students in my populations, that there are no differences in the 'difficulty' of the subjects or their examination artefacts, the findings merely reflect the different ways by which the different populations have interacted with the assessments. If that is so, it does not explain the significant similar change in how they react from 1994 to 1995 with physics and chemistry becoming respectively 'easier' and 'harder' when examined by two different examining groups with their different assessment artefacts for populations that are differently constituted, for example in terms of student variables such as motivation and in their geographical locations and examination centres. More questions seem to be raised than answered by this analysis and the value of engaging teachers in discussing these issues for illumination purposes is indicated.

Correlation analysis

Spearman correlation coefficients and their significance are summarized in Table 4.11. As for the WJEC populations, a positive correlation between the SEG 1994 population's GCSE science grades occurs significantly (P <0.001) for biology and chemistry, biology and physics, and chemistry and physics. Obtaining a high grade on one science is apparently predictive of obtaining a high grade in another science for both examining groups. One interpretation is that even though WJEC and SEG have different science examinations, these evoke interactions with their populations that produce positively skewed grade distributions. It is tempting to speculate from a constructivist view that this is because the same types of skills are required and / or the same types of items (questions) are present in the different groups' science examinations.

Table 4.11 Spearman Correlation Coefficients Between Biology, Chemistry and Physics Grades – SEG

	Chemistry	Physics
Biology		
1994	0.76	0.75
1995	0.64	0.59
Mean	0.70	0.67
Chemistry		
1994		0.79
1995		0.65
Mean		0.72

N for 1994 = 1001
1995 = 1759 for biology and chemistry, 1758 for physics pairings
All correlation coefficient values are significant at the 0.1% level

Again as for the WJEC populations, the chemistry and physics grades are the most positively correlated of the three subject combinations, followed by biology and chemistry grades which in turn are slightly more positively correlated than the biology and physics pairing. The same holds for the 1995 population but the biology and chemistry pairing's value is closer to that for chemistry and physics than that for biology and physics. The similarity in the order of the subjects' positive correlation values *suggests* that there might be factors influencing students' performances in these subjects or in the marking / grading processes that transcend examining groups. From a cognitive constructivists view, one would expect this trend if the skills required by the chemistry and physics examinations are more similar than those required by physics and biology or chemistry and biology.

Kappa analysis

Both the 1994 and 1995 primary datasets required amendment to secure matched grades for kappa calculations. This procedure reduced the 1994 primary dataset from 1001 to 998 students and similarly the 1995 population from 1761 to 1758. Table 4.12 shows the resulting kappa values.

Table 4.12 Kappa Values for Biology, Chemistry and Physics - SEG

	Chemistry	Physics
Biology		
1994	0.305	0.280
1995	0.323	0.282
Mean	0.314	0.281
Chemistry		
1994		0.263
1995		0.297
Mean		0.280

All Kappa values are significant to 0.1%

As for all four WJEC populations, there is 'fair' (Landis and Koch, 1977) agreement between students' achieved grades for all of the science subject pairings for both the 1994 and 1995 SEG populations. The biology and chemistry pairing has the highest level of agreement for both SEG populations. The lowest level of agreement is shown by chemistry and physics in 1994 and by biology and physics in 1995. However, there are only small differences in the kappa values for all of the different subject pairings and in particular for the physics pairings with biology and chemistry. Overall, the mean kappa values indicate that it is more likely for students to obtain the same grade in biology and chemistry than in the other science subject pairings in the 1994 and 1995 examinations under scrutiny. This is in contrast to the WJEC findings where most agreement was shown most often for the physics / chemistry pairing. Other than all science subjects showing 'fair' agreement between students' grades there are no other patterns in kappa values that hold across the two examining groups. Overall, the analyses indicates that there is only a 'fair' (ibid.) chance of obtaining identical grades in the science subjects at Triple Award GCSE. The notion of subject 'gradeness', in the sense that if a student obtains a particular grade in one science subject it is predictive of the same grade in another, is not strongly supported by the findings across both WJEC and SEG. Arguably the findings support the view that students' interactions with examination artefacts vary in their nature from student to student and thus population to population. One would not expect identical grades to be obtained by the same population across different subjects even if these have similar skill demands as there is still the potential for

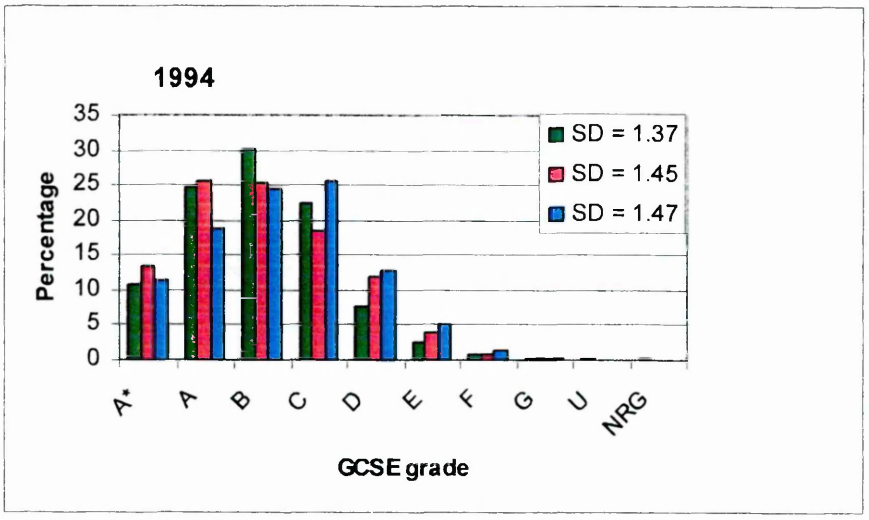
differences in interaction with items (questions), for example in what skills the students actually use and / or the contexts used. The differences in the kappa values for WJEC and SEG, and in particular for the physics and chemistry pairing, also illustrate the sensitivity of the meaning attributed to comparability. The comparability outcomes can vary for the same subjects for the same periods of time according to the source of data even when the same statistical treatment is used.

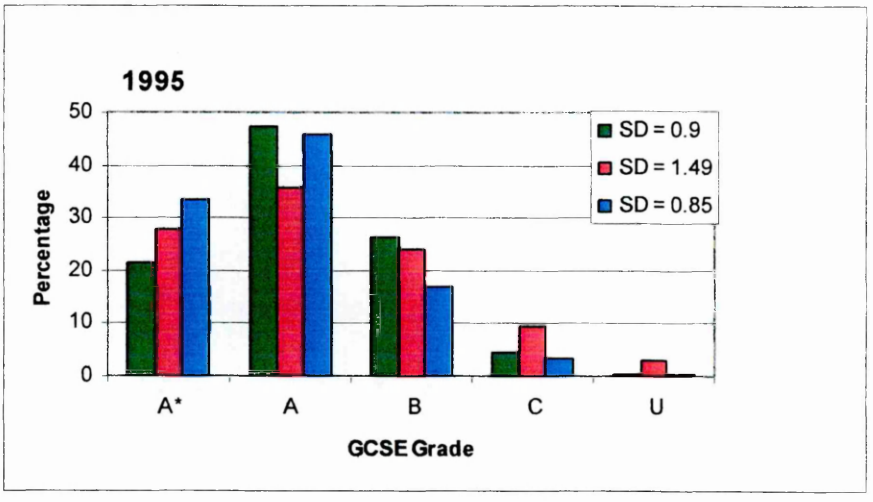
Descriptive statistics

Appendix 6 shows the frequency, percentage, cumulative percentage, means and standard deviations for the students achieving each GCSE grade in biology, chemistry and physics. The bar charts in Figure 4.10 illustrate some of this information. The standard deviation values, although similar, reveal a slightly larger distribution of the 1994 population's achieved grades for physics (1.47) than chemistry (1.45), with the smallest distribution being for biology (1.37). This order replicates that for the degree of severity of grading from most to least severe based on achieved means. The cumulative percentages indicate a significantly lower attainment of grades A and B in physics than in biology and chemistry. All of the achieved science subject grades are positively skewed, with physics being the least positively skewed, followed by chemistry then biology.

The standard deviation values for the 1995 population show a significantly larger distribution of the achieved grades for chemistry (1.49), than for either biology (0.90) or physics (0.85). This agrees with the order of degree of grading severity (most to least severe) identified from the achieved means. The pattern of SEG's standard deviation values for biology and physics decreasing from 1994 to 1995 reflects the same trend for all of the science subjects for WJEC from 1994 to 1995. This again highlights the examination of new syllabuses by examining groups in 1995 being associated with changes in patterns of assessment outcomes. Chemistry's position as the most severely graded science subject is not reflected in a lower percentage of grade A*s (biology has proportionally more) but rather in fewer grade Bs and more of the lower grade C and unclassified. If students failed to achieve grades A* - C in this particular SEG examination (high option, 1995), then the policy was to award them grade U and none of the grades D - G inclusive (SEG, 1996). Chemistry appears to have a

Figure 4.6 SEG GCSE Biology, Chemistry and Physics Grade Distributions





disproportionate number of students who fall into this category and therefore appear to have been entered for an inappropriate examination. Here the technical approach illuminates the importance of teachers' decisions on students' tier entry for students' achievements and the need to explore this with teachers.

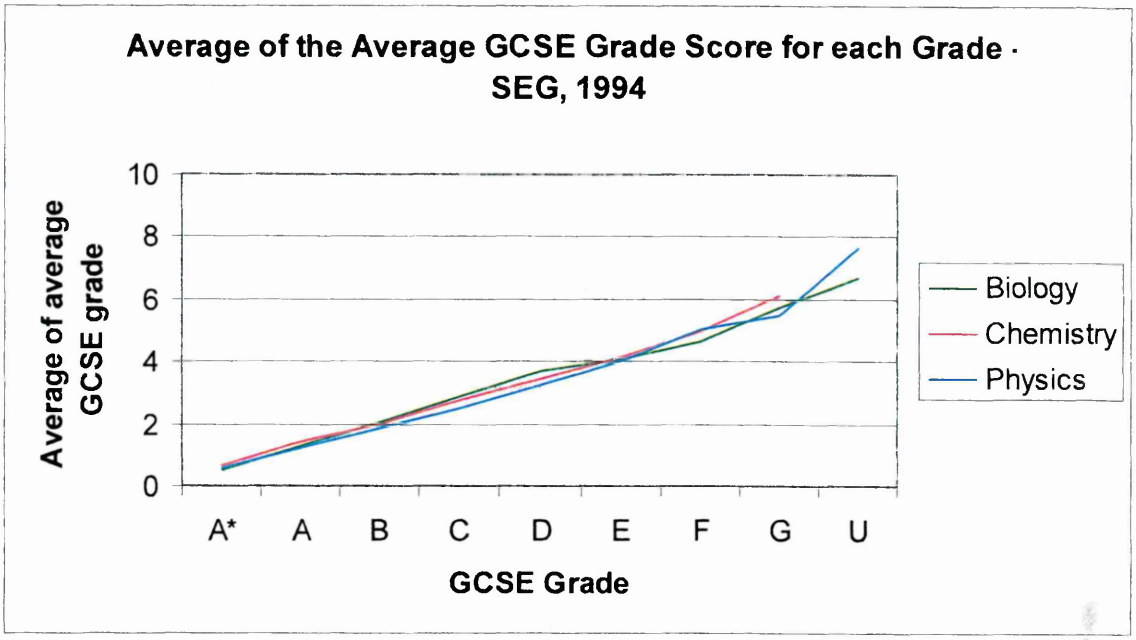
4.8.2 Are there relationships between students' performances in biology, chemistry, physics GCSE examinations and their average GCSE grade scores?

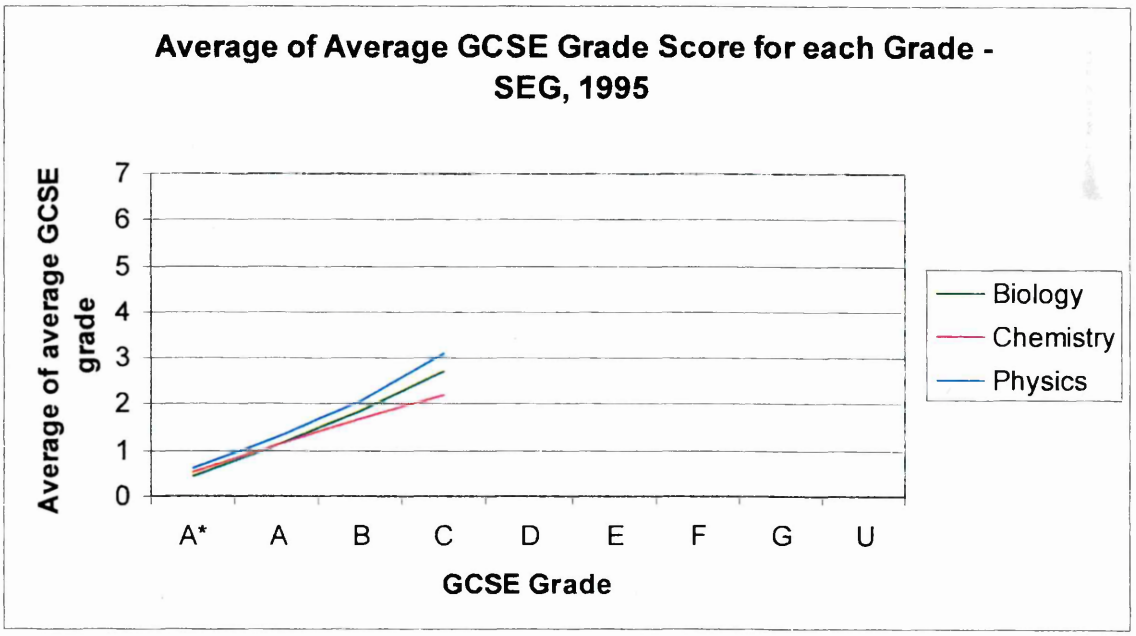
Graphical analysis

The average of the average grade scores for each grade (A*-U) in each of the SEG 1994 population's biology, chemistry and physics results was produced from the associated primary data set. The same treatment was conducted on the 1995 population for its associated grades A* to C and grade U. Figure 4.7 illustrates the analysis outcomes with line charts (graphs). The gradients of the lines in the line chart give an indication of the relative difficulty of the subjects biology, chemistry and physics for the study's populations with the generalization that the steeper the gradient and / or the higher up the y axis the line appears, the 'easier' (less severely graded) the subject.

There is no consistency in the science subjects' slope difference in the 1994 graph making it difficult to comment on the subjects' relative severity of grading, which in my view even if one supports the technical approach to comparability, limits the usefulness of this treatment in technical comparability investigations. There *is* an overall tendency for physics to be the most 'difficult' (severely graded) science subject, which concurs with the finding from the subject-pair analyses. Overall, the 1994 population has done rather less well in the science subjects than in their GCSEs as a whole. For example, students achieving grades C, D, E or F in the study's GCSE science examinations have average GCSE grade scores which are respectively better. For these students, the associated science subjects would appear to have been 'harder' than most of their other GCSE subjects as a whole. The science subjects' line slopes are more consistently related to each other for the 1995 population than for the 1994 population. Overall, physics is 'easier' (less severely graded) than either biology or chemistry for the 1995 population under scrutiny. Chemistry is *overall* 'harder' (more severely graded) than biology. These findings concur with those from the subject-pair analyses. As for the 1994

Figure 4.7 Graphical Analysis: SEG 1994 and 1995





population, overall the 1995 population has a tendency to achieve lower science GCSE grades than is the case for the majority of their other GCSE subjects. Furthermore, overall, students tend to perform less well in their biology, chemistry and physics examinations than in their other GCSE subjects in *all* WJEC *and* SEG populations. From a technical point of view this implies that the science subjects were more severely graded than other GCSE subjects and in that sense 'gradeness' is challenged with it being apparently harder to get high grades in science subjects than others.

4.8.3 Are there relationships between students' sex and their achieved SEG GCSE biology, chemistry, physics grades and average GCSE grade scores?

Inferential statistics

Although both of the SEG populations of this study contain a majority of boys, the 1995 population is dominated by more boys (M = 1179 ≡ 67%: F = 580) than that for 1994 (M = 545 ≡ 54.4%: F = 456).

For the 1994 population there are no significant differences in the boys' and girls' sub-groups' achieved physics GCSE grades. However, overall, girls achieved significantly ($P = 0.000$) better grades in biology and chemistry, and performed significantly ($P = 0.000$) better in their average GCSE grade scores than boys. The same pattern holds for the 1995 population but in respect of biology and average GCSE grade score with a lower level of significance. Thus in terms of average GCSE grade scores, girls out-perform boys in *all* WJEC and SEG populations at a statistically significant level. Arguably this outcome could be expected as my populations are '*restricted*' samples (Willingham and Cole (1997, p98) and the minority group of girls would be expected (ibid.) to outperform the majority group of boys. The populations only consist of students who have been entered for all three Triple Award science subjects in the same tier giving access to top grades. The girls who fall into this category could be said to be more highly motivated than boys in order to sustain study in physics, which is popularly regarded as a boys' subject. Comparing sex sub-group performances across the years of available data for both WJEC and SEG indicates girls out-performing boys in chemistry at a significant level in 1994 but with no such similarity in 1995. Interestingly, the trend in girls under-performing boys in WJEC physics, despite their overall better performance in their GCSEs as measured by their average GCSE grade scores, is not shown for the SEG sub-groups. This illustrates

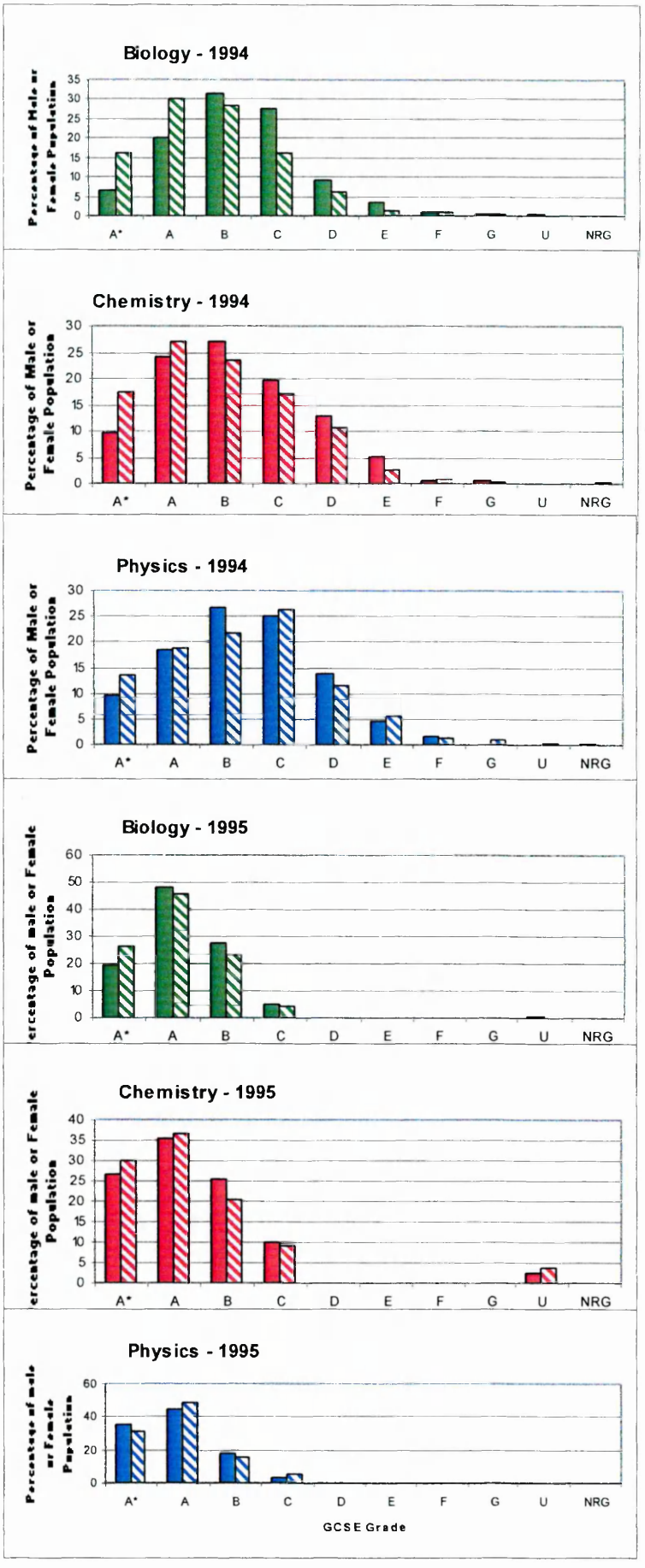
the 'danger' of drawing conclusions from the technical findings in comparability investigations. It is not valid to say that boys out-perform girls in GCSE Triple Award physics - my findings are context specific – at best it could only be said of my selected sub-group samples in my chosen examinations using my chosen statistical treatments. Importantly, the difference in findings for girls in WJEC and SEG physics examinations reinforces the view that the girls' poorer performance than boys in WJEC physics is not due to physics *per se* and supports the notion that the WJEC findings are due to differences in the sub-groups' interactions with those particular examinations. This finding strengthens my argument about students' different types of interactions with assessment artefacts influencing their achieved grades as vital considerations when comparing grade distributions and the shortcomings of the traditional technical approach which largely ignores them.

Descriptive Statistics

Continuing with the technical approach, descriptive statistics were used to explore the results of the above inferential tests. Cross-tabulations of sex with each of the variables biology, chemistry and physics grades in the form of frequencies for each grade category are illustrated in Figure 4.8. As for the WJEC populations of this study, the analysis was based on percentages of each sex's sub-group rather than the population as a whole to increase the validity of comparing the different sex's GCSE achievements.

For the 1994 SEG sex sub-groups, girls tend to achieve proportionally more A* and A grades than boys in all three science subjects and particularly so in biology and chemistry. This pattern reflects girls overall better performance than boys in achieved mean values in biology and chemistry. Overall in terms of mean values, boys perform better than girls in physics and yet girls achieve proportionally more of the top grades (A* and A) in this subject. This illustrates the problem of treating sub-groups as homogenous in the technical approach – here one could say that girls have out-performed boys at grades A* and A in physics despite boys apparent overall better performance in this subject. Similarly, the SEG 1995 counterparts show boys achieving proportionally more of the top A grades in biology despite girls overall performing significantly better in terms of achieved mean values in this subject. This phenomenon is also observed within the WJEC sex sub-groups. For example,

Figure 4.8 Distribution of Boys' and Girls' Grades – 1994 and 1995, SEG
Block of colour = Boys; Hatched colour = Girls



overall the girls mean value is significantly better than boys in WJEC chemistry in 1993 and yet boys achieved proportionally more of the top grade B (grades ran from A-G).

Continuing with this theme, the trends in science subject grading difficulty show that the SEG 1994 population subsumes a gender sub-group difference, as was observed for the WJEC 1993, 1994 and 1995(02) populations. Based on the achieved mean values, the 1994 boys sub-group's performance is lowest in chemistry whereas this is the case for biology for their female counterparts. However, both the boys and girls sub-groups found physics to be the 'hardest' of the science subjects. No sex differences in subject 'difficulty' appear to be subsumed within the SEG 1995 population. Table 4.13 also indicates that a change in the pattern of science subject severity occurs in the 1995 examinations, at which point chemistry ceases to be the most leniently graded science subject for *both* sub-groups – this coincides with all GCSE examining groups examining newly introduced science syllabuses for the first time.

Table 4.13 Differential Gender Performance, SEG

Population	**1994**			**1995**		
Science Subject	**M+F**	**M**	**F**	**M+F**	**M**	**F**
Most severely graded	**P**	**P**	**P**	**C**	**C**	**C**
Least severely graded	**B/ C ***	**C**	**B**	**P**	**P**	**P**

B = Biology; C = Chemistry; P = Physics

B/C* denotes Biology and Chemistry as being the same in terms of severity of grading

This finding mirrors that for the WJEC populations and strengthens the argument that changes in the GCSE arrangements consequent on incorporating the national curriculum in its syllabuses had an effect on the relative severity of grading. My view is that the severity of grading did not change *per se*, rather the subjects redefinition caused a change in how the students' interacted with the associated assessment artefacts.

4.9 Summary and ways forward

4.9.1 What has and hasn't the technical investigation bought me?

The technical findings illuminate the notion of 'gradeness' by identifying aspects of students' relative performances in different GCSE subjects. In that sense my first two research aims have been

addressed as I have illuminated the anecdotal evidence provided by teachers concerning comparability of grading described in Chapter 1. To summarise, my technical investigation's findings may be interpreted as showing that students tended to do less well in their science subjects than in their other GCSE subjects. This pattern held across both WJEC and SEG populations and could be the equivalent difference of a whole grade. The attainment of high grades in biology, chemistry and physics were all individually associated with an overall high performance in GCSE subjects with correlations ranging from 'moderately' to 'strongly' positive. Again this trend held across WJEC and SEG populations. No one particular science subject was identified as being most severely or leniently graded for *all* of the populations investigated. In that sense the two concerns expressed by separate groups of science teachers reported in Chapter 1 are not substantiated by the technical findings because neither chemistry nor biology are found to be consistently the most severely graded of the science subjects across the examination sessions in my investigation. The apparent fluctuations in the science subjects' severity of grading across time counter any notion of a subject being inherently more 'difficult' than the others and also challenge the ways by which comparability data may be interpreted across years. The apparent change in the science subjects' severity of grading for both WJEC and SEG consequent on 1995's first examination of new syllabuses based on the national curriculum. highlights the potential for curriculum-examination interactions. For both examining groups physics became significantly less severely graded and chemistry became more severely graded from 1995, whilst biology became slightly more severely graded, thus supporting the biology teachers' views regarding their subject reported in Chapter 1. Additionally, the science subjects' correlation and standard deviation values for both examining groups and WJEC's examination paper cognitive skill demands all change from previous patterns from 1995. It is possible that the teachers' views are the result of their picking up on the *effects* of curriculum –examination interactions consequent on the 1995 GCSE syllabus changes.

Overall a high performance in one science subject was moderately associated with a high performance in another science subject. For both WJEC and SEG populations, physics and chemistry were the most positively correlated of the science subject pairings followed by biology and chemistry,

which were slightly more positively correlated than biology and physics. Students were more likely to obtain identical grades in physics and chemistry and least likely to in physics and biology. Although both English and mathematics correlated positively and significantly with the individual science subjects, the correlations were overall more positive for mathematics than English. The order of positive correlation between the students' achieved mathematics grades and science grades was in the order biology (least positive), chemistry, physics (most positive) and this was consistently shown across the years of the study. Thus although the severity of grading of the science subjects varies across the years of the study, the correlation findings for the physics and chemistry pairing, and mathematics and the science subjects is relatively stable across all populations, arguably indicating an influence which elicits similar student-examination interactions across these subject pairs. I interpret this as evidence of similar skills being required for students to interact with the physics and chemistry examinations and the mathematics and science subject examinations in my study, which as an influence on examination performance appears to hold as significant in the presence of any other influences. In that sense the notion of 'gradeness' is challenged. If GCSE subjects vary in their skill requirements, the grades can only still claim to have common currency across these subjects if the grade awarding process facilitates this as claimed by examining groups. As discussed in Chapters 2 and 3, examiners / grade awarders function as a 'guild of professionals' (Sadler, 1987) with an understanding of their particular subject and how examination performance equates to specific grades but not with such an understanding for other subjects. Arguably, a grade A in physics cannot have the same currency as a grade A in biology when they are associated with different skill demands.

For the girls' sub-groups, physics was significantly the most severely graded science subject; this was not the case for the boys' sub-groups. This trend is seen more often for the WJEC than the SEG sub-groups. Biology was significantly the most severely graded of the sciences for three out of the four WJEC boys' sub-groups. This trend was not observed for any of the SEG boys' sub-groups. Overall for both WJEC and SEG populations, there was a trend for boys to outperform girls in physics and for girls to outperform boys in biology and chemistry. Where boys overall outperform girls in a science subject, girls may still achieve proportionally more of the higher grades in the subject than the

boys. However, given the caveats covered in Chapters 3 and 4, throughout my technical investigation I have been constantly qualifying my technical findings to make valid interpretations, acknowledging they are specific to my selected populations and sub-groups and emphasizing their limited dependability. For example the sex sub-group analytical techniques treat each sub-group as homogeneous whereas the group members have their own identities. Trends for the sub-groups may not be attributable to all members of that group. There may be great variation *within* each sub-group and there may be considerable overlap *between* the sub-groups.

This highlights a fundamental concern for any comparability study. The way that differences in performance of girls and boys as groups - and students *per se* - are understood is determined by understandings about the nature of learners. As Murphy and Whitelegg note (p. 46, 2006), if assessment tasks are seen as neutral devices as I believe they are in the technical approach to comparability, that all students understand in the same way, then differences in performance can be attributed to either innate differences or differences in opportunities to learn. However, if a constructivist view is taken, account needs to be taken of learners actively constructing meaning and all assessment tasks like learning tasks are interpreted by students. Their interpretations will depend on what they bring to the task, for example their experiences and motivation. Consequently, differences in performance *may* reflect differences in achievement and opportunities to learn – but in my view may also indicate that students are responding to different tasks than those intended and what is being assessed is not stable across students or sub-groups of students. This view requires account to be taken of how an assessment task might influence students' understanding of its requirements. Influences on student / assessment task interaction as discussed in Chapter 3, including item (question) format, content, context, and assessment format, for example coursework and examinations, need to be taken into account when comparing students' assessment performances. Arguably if girls as a group were less familiar with the contexts of items used by examining groups in physics than biology, items that more often linked to boys' than girls' interests and opportunities to learn outside of school, then the trend of girls underperforming boys in physics would be anticipated across populations and across examining groups. That this happens to a point (all WJEC populations) provides evidence of sex

effects but because the underperformance of girls in physics does not hold for the SEG populations, there is also the possibility of this being disrupted. The same argument may be applied to the underperformance of boys relative to girls in biology in some of the populations in my study. It is tempting to suggest that in the examinations considered in my study SEG provided physics examination papers that more closely aligned to boys *and* girls interests and opportunities to learn outside of school than WJEC.

The validity of comparing my populations' performances across time is challenged not least by the differences in the nature of the populations as described in this Chapter. Although the populations are similar in several respects, they *do* differ from each other and in particular in respect of the relative proportions of students based in 11-16 and 11-18 type schools. Again as shown in this Chapter there have been significant changes in the nature and weighting of coursework in each of the three sciences, again challenging the validity of comparing performances across time. For the 1993 population there was also a difference in the nature *and* weighting of their physics coursework as opposed to that for their biology and chemistry, which challenges the validity of comparing performances across the three sciences for this year, although examining groups claim (WJEC, 1995) that the processes of moderation and grade awarding should overcome this challenge.

I have commented on the relative usefulness of the various statistical treatments for investigating comparability as their findings have been presented. For example graphical analysis of achieved GCSE grades against the average of the average GCSE grade is not useful when the sloping lines cross each other as for WJEC 1995 (02) and SEG 1994. Kappa analysis appears better suited to larger than smaller populations to ensure that all available grades have been achieved by students on all considered subjects; I had to substantially amend my datasets to take this issue into account and this impacted on the usefulness of the findings from this technical approach. The descriptive statistics proved useful in clarifying the profiles of achievement; I can say how dispersed the students' achieved grades were about the pass grades and describe the relative proportions of high grades achieved by the sex sub-groups in different subjects, all of which enriches what can be revealed about comparability. For example girls can obtain proportionally more of the high A grades than boys in a subject even

when their group's mean grade is lower than that of the boys' group, as seen for the WJEC physics 1994 and 1995(03) sub-groups. Obtaining complete sets of data also proved problematic in the case of tracing students' 'missing grades' for mathematics and English. Consequently, the findings from the datasets not having all of the students' mathematics and English achieved grades would not be viewed by examining groups as such valid indicators of my populations' achievements as those for students' biology, chemistry and physics achieved grades.

4.9.2 Reflection: ways forward and a shift in my theoretical position

This Chapter and Chapter 3 have consistently shown that technically, investigating comparability is hugely problematic. Despite the numerous technical treatments that can be used, it is simply not possible to control for the varied influences on examination performances that can affect the validity of performance comparisons. In my view this is because an examination performance cannot be treated as an entity divorced from what I see as 'unquantifiable' forces that shape it, for example student motivation. As I have been interpreting my technical findings, I have been increasingly drawn to consider social aspects of the assessment process and to reflect on my own experiences as a GCSE examiner and teacher entering my students for GCSE examinations. The rest of this section follows this line of my thinking and recollections.

In my technical investigation I attempted to increase the validity of comparing students' science performances by taking populations consisting of students entered for the same tier of available grades in all three science subjects. The accuracy with which teachers can predict students' national examination grades has been well-investigated (Murphy, 1979, 1981; Petch, 1964; Sowell, 1970) and consistently shows a reasonably high level of agreement between teachers' predictions and actual awarded grades, although arguably this could be self-fulfilling. Rather less well researched is the accuracy with which teachers enter students for appropriate tiers in differentiated examinations (Good and Cresswell, 1988d). Good and Cresswell (1988d) report that there may be a considerable number of inappropriate entries to GCSE examinations that use differentiated papers and Gillborn and Youdell (1998) show that this is particularly true for Black students. Stobart, Elwood and Quinlan (1992) identified a tendency for girls to be entered for tiers in mathematics GCSE examinations that

were not commensurate with girls' mathematical ability. Intermediate rather than higher tiers tended to be allocated as a 'safety first' approach, reinforcing a perception that girls were less confident of succeeding than boys.

From the literature and my experiences as a teacher and GCSE examiner I know that inappropriate tier entries occur. I have worked in schools where teachers 'play safe' and tend to enter students for tiers in which grade C is the top available grade, fearing that entry to a higher tier risks students succumbing to the 'floor' effect and failing to achieve a grade. I accept the possibility that some students in my populations may have been entered for inappropriate tiers of examination papers. For example, in one particular subject a student may be capable of achieving grade A but has been entered for a tier that only gives access to grade C. In other subjects this condition may not apply. Comparisons of such performances might lead one to suggest that a student is only *capable* of grade C in one subject but a higher grade in another subject. This condition could apply for an unknown number of such students and questions the validity of comparing students' performances even when they are selected, as I have, from the same tier of subject examination papers.

If I am to understand better the notion of 'gradeness' and so continue my exploration of examination comparability, the way forward does not seem to lie within a technical but rather a social approach. My technical findings, despite my caveats about their validity, *have* illuminated 'gradeness'. I could use them as a resource for my continuing exploration as exemplified by the following. The significant positive correlation between physics and chemistry and the least positive correlation between biology and physics for both WJEC and SEG datasets and the consistently most positive correlation of mathematics with physics could all be explored for associations with assessment artefact issues. Another example comes from my finding that there appeared to be a change in the severity of grading of the Triple Award science subjects from 1995 with the first WJEC and SEG examination groups' examinations set on the national curriculum. Simultaneously, physics became significantly 'easier' and chemistry became significantly 'harder' and this was associated with significant changes in the individual science subject cognitive skill demands of the WJEC examination papers. Additionally, the 1995 correlation coefficient values for the biology, chemistry and physics

pairings showed more dissimilarity with each other as well as with the values for the 1993 and 1994 populations. The standard deviation values also became more similar for the three science subjects for the 1995 examinations compared with those of 1993 and 1994. All of which indicates the potential usefulness of exploring the nature of the *actions* causing these changes. They could be due to assessment artefact changes. For example do the syllabus contents reflect changes in how physics and chemistry are variously defined – and can any such changes be aligned with socio-economic and political pressures, as could be claimed for physics if it were being made 'easier' to stem the decreasing numbers of students studying it post 16?

I choose not to focus primarily on assessment artefacts or on examining group policies and actions for associations with my technical findings as the literature and my teaching and examining experiences have made me more interested in the assessment process as it is played out in schools. The genesis of this research lay in anecdotal evidence from teachers – *their* concerns about differences in the severity of grading of biology, chemistry and physics. As an ex–teacher I remain in contact with many teaching colleagues who would be willing to give me their time for extending my investigation of comparability. Engaging with teachers would enable me to explore their beliefs and practices in relation to assessment – if and how they mediate assessment and the relationship this has with comparability. For example, I could explore my finding of mathematics being more positively correlated with physics than chemistry or biology by obtaining their views on the mathematical demands of physics and how these views play out in their actions in relation to students' assessment. I could explore teachers' views of what is important in allocating their students to specific tiers of GCSE entry and in turn how this impacts on students' performances. Teachers' views on the impact of examining new syllabuses in 1995 could illuminate my findings of a significant change in the associated examination cognitive skill demands and the simultaneous changes in severity of grading of physics and chemistry. My kappa findings indicate a fair measure of 'subject gradeness' – that there is fair agreement of obtaining the same grades in the different science subjects and that this is most likely for physics and chemistry. I could extend my exploration of 'subject gradeness' by identifying whether teachers support it as a notion, the ways in which they understand it and how it plays out in

their practice in relation to their students' assessment. In doing this, I would also be exploring teachers' views of the relative difficulty of biology, chemistry and physics and relating these to my severity of grading findings. Furthermore, my technical findings have shown that different meanings can be attributed to comparability. What is used as a descriptor of comparability can vary according to the statistical treatment used and the validity attributed to the treatment by various persons. For example subject pair analysis is an acceptable measure of comparability by examining groups; the percentages of achieved grades for sub-groups widely used by AQA is not an acceptable measure of comparability for some researchers such as Gorard (Gorard *et al.*, 1999). I could explore teachers' understanding of comparability and seek evidence of how this relates to their practice and affects their students. By engaging with teachers I can extend my understanding not just of comparability but of the social nature of the assessment process itself.

It is at this point that I came to revise my theoretical position. Reflecting on my time as a teacher I recalled what it was like to teach – to function within the classroom, within a school. I did not recall how I prepared and entered my students for GCSE science examinations being consciously influenced by 'this' school policy or by 'that' out-of-school issue. Rather I recalled a myriad of interactions with different types of staff, parents and students and my teaching and assessment practices emerging from this amorphous milieu. I therefore looked at other theoretical positions appropriate to my intention to explore teachers' beliefs and actions in their classrooms for extending my understanding of examination comparability. The next Chapter describes this journey towards my adopting a sociocultural approach for the qualitative dimension of my research.

CHAPTER 5

The qualitative investigation: theoretical position and research design

5.1 Locating my theoretical position

On writing about examination techniques and issues of validity and effects on students' performance Elwood notes that the measurement of students' performance is not an exact science, *'but a process that is underpinned by subjective judgments and value-laden choices, albeit made with fairness and validity in mind'* (2001, p. 100). Examining and assessing processes and the means by which final results are determined *are* social constructs (Cresswell, 1996). Assessment is not an activity that is *done* to students, but is fulfilled through a process of *social interaction* in which the practices of the participants have a critical effect on the outcome (Pryor and Torrance, 2000). Assessment outcomes are actively produced rather than revealed and displayed by the assessment process (ibid.). Each participant, variously a teacher, a student and an examining group employee, brings to the assessment process their own understandings of a myriad of issues. Pryor and Torrance (2000) argue that these understandings are then subject to change as a result of the inferences that are made during the interactions of the participants. I wish to illuminate this view of assessment as a social dynamic process with my focus on teachers' practices and beliefs, and specifically in relation to their tier entry decisions and how this relates to comparability.

Research into teacher education demonstrates that knowledge of self, i.e. identity, is a crucial element in the way that teachers construe and construct the nature of their work (Kelchtermans and Vandenberghe, 1994). Events and experiences in teachers' personal lives are also shown to be closely linked to their professional performances (Ball and Goodson, 1985; Goodson and Hargreaves, 1996; Acker, 1999). Researchers such as Nias (1989, 1996), Hargreaves (1994) and Sumison (2002) have all shown that teacher identities are not only constructed from technical and emotional aspects of teaching such as classroom management, subject knowledge and student test results, and their personal lives, but also *'as the result of an interaction between the personal experiences of teachers and the social, cultural and institutional environment in which they function on a daily basis'* (Sleegers and Kelchtermans, 1999, p. 579). As James-Wilson (2001) notes:

> *The ways in which teachers form their professional identities are influenced by both how they feel about themselves and how they feel about their students. This professional identity helps them to position or situate themselves in relation to their students and to make appropriate and effective adjustments to their practice and their beliefs about, and engagement with, students.*
>
> (James-Wilson, 2001, p. 29)

Therefore, when exploring teachers' practices and beliefs in relation to assessment I need to take account of the social, cultural and institutional environments that constitute their schools.

In the 1960s it was still unusual for a woman to study chemistry at university as I did. As a chemistry teacher in the late 1960s it was also not uncommon for me to have my professional identity questioned with comments such as *'isn't it hard being a woman teaching a man's subject?'*. As I have found from my own experiences, personal and professional identities are interrelated (Day *et al.*, 2006). Both evolve over time (ibid.), although researchers disagree on how stable teachers' identities are and the degree of plurality that exists within them (Beijaard, 1995; Cooper and Olsen, 1996; Reynolds, 1996). The literature cited so far suggests that identities are a changing meld of personal history, culture, social influence and institutional values which may change according to circumstance. As Day (2006) notes, it is the combination of the variations in teachers' work *and* lives, in addition to the strategies adopted by teachers to deal with any arising tensions between them, that determine the individual identities for each teacher and which in turn may have a direct or indirect positive or negative influence on students. For example, the views I encountered as a chemistry student and teacher make me particularly encouraging of girls wishing to pursue mathematics or science studies. Therefore, as I access teachers' practices and beliefs in relation to assessment, I will need to be aware of a myriad of interrelated personal and professional issues. For example, research (Sikes *et al.*, 1991) shows that for secondary school teachers, subject and its status are related closely to identity, and more so than for primary school teachers. To take just one example from my technical findings on boys' and girls' performances, I could usefully explore teachers' views of their subject for connections with their assessment practices for entering boys and girls for different tiers of GCSE examination papers.

As a consequence of the above, I sought a theoretical position and methodological approach that would anticipate the mediation of social practices and structures by individuals and would consider the interactions between teachers' actions and assessment structures. This was to get at what might lie behind grade distributions that raises further questions about claims of, and meanings attributed to, notions of examination comparability and help me better understand 'gradeness'. I was initially drawn to a humanistic paradigm that is based on the belief that human behaviour can not be understood without reference to the meanings and purposes attached by human actors to their activities (Guba and Lincoln, 1998) because it reflected my teaching experiences in that I had not been able to understand the behaviour of my students without referring to *their* understandings. Ontologically, a humanistic paradigm is underpinned by a belief that realities may be understood in the form of *'multiple, intangible mental constructions, socially and experientially based, local and specific in nature'* (Guba and Lincoln, 1994, p. 110). Language is seen as shared tools which have meanings in social interaction and in specific contexts. I have found throughout Chapters 2 to 5 that the discourse on assessment and its meanings have differed between groups of people such as examining group employees and educationists. Therefore I have come to view truth as a matter of the best informed construction on which there is consensus at a particular time (ibid.) and as socio-culturally relative. This view concurs with an epistemological position that is transactional and subjectivist (ibid.). So meanings are shaped by language and other social processes; knowledge and learning occur through the shared activities of people. Because of this shift in my thinking, I was drawn to a sociocultural approach to understanding.

Wertsch describes a socio-cultural approach in the following way: *'the basic goal of a socio-cultural approach to mind is to create an account of human mental processes that recognises the essential relationship between these processes and their cultural, historical and institutional settings'* (1991, p. 6). As Murphy and Ivinson (2004) note, when this approach is applied to assessment it is no longer possible to assume that similar awarded grade distributions for different examinations indicate the validity of the assessments. They state that: *'Messick in no way refers to himself as a socio-culturalist, nevertheless his challenge to traditional 'types' of validity takes account of socio-cultural influences and the social nature of assessment as a process and a*

product', (2004, p. 371). Messick, cited by Murphy and Invinson (2004), argues for an over-arching concept of validity, which he calls construct validity and describes as:

> *'... a sine qua non in the validation not only of test interpretation but also of test use , in the sense that relevance and utility as well as appropriateness of test use depend, or should depend, on score meaning. To act otherwise is not just dubious but dangerous. Using test scores that 'work' in practice without some understanding of what they mean is like using a drug that works without knowing its properties and reactions'.*
>
> Messick, 1989, p. 162

Although there are major differences amongst authors in the way that Vygotsky's ideas on human mental processes should be understood and applied, there is a shared conviction (Minick, 2005, p. 53) that these ideas constitute a conceptual framework that overcomes many limitations of other attempts to represent the relationship between the social and the individual in psychological development. Vygotsky's ideas specify individual mental functions as developing from socio-cultural processes - that mental functioning reflects and embodies its historical, institutional and cultural setting (Vygotsky, 1978). The individual participates in social activity mediated by speech and by psychological tools that others use to influence her behaviour and that she uses to influence the behaviour of others. According to Vygotsky (1981, p. 137) the following can serve as psychological tools: language; various systems for counting; mnemonic techniques; algebraic symbol systems; works of art; writing; schemes, diagrams and maps. In all cases, these 'tools' are *'mediational means that are the products of sociocultural evolution and are appropriated by groups or individuals as they carry out mental functioning'* (Wertsch and Tulviste, 2005). Subsequently, the individual *'begins to apply to herself the same forms of behaviour that were initially applied to her by others'* (Vygotsky,1960, p. 192.). In this way, the organisation and means of social activity are taken over entirely by the individual and appropriated, leading to the development of mediated, voluntary, historically developed mental functions that, as Minick (2005, p. 38) describes, *'are based on stimulus-response components but cannot be reduced to them'.* Vygotsky referred to these psychological processes as 'higher mental functions' and formulated a general principle underlying their development as:

'Any higher mental function was external [and] *social before it was internal. It was once a social relationship between two people ... We can formulate the general genetic law of cultural development in the following way: Any function in a child's cultural development appears twice or on two planes ... It appears first between people as an intermental category, and then within the child as an intramental category. This is equally true of voluntary attention, logical memory, the formation of concepts, and the development of will.'*

(Vygotsky, 1960, p. 197-198)

On so describing the mental activity of a person as a function of social interaction he refers to children, but it applies to how *all* social actors make meaning. Research traditions such as social learning theory or cognitive anthropology are based on the concept that there are important mechanisms of learning and development that are inherently social. Minick (2005, p. 38) sees Vygotsky as taking this further by linking the social not only with unique mechanisms of psychological development such as social interaction and appropriation but with types of mental processes that are themselves inherently social, specifically the higher mental functions. Human consciousness and behaviour become aspects of an *integral* system – mental functions develop not merely through an individual's experience in social interaction but through the transformation of social behaviour from the intermental to the intramental plane.

This sociocultural approach to learning emphasises classroom practices as being situated and mediated by processes beyond school. This emphasis reflects my view of learning based on my experiences as a teacher. However, Vygotsky specified little in terms of how his approach applies in concrete settings (Wertsch, 1991). The notion of a *community of practice* as developed by Lave and Wenger (1991), Rogoff (1995) and Wenger (1998) has been used relatively widely to characterize teaching communities and allows a Vygotskian sociocultural approach to be applied to the schools and teachers in my study. The three interrelated dimensions that Wenger uses to characterize a community of practice are found in the cases I chose to investigate. First, the teachers I wish to engage with are occupied in a joint enterprise, for example preparing their students for Triple Award Science GCSE examinations. Second, they are in mutual relationships that encompass norms of participation, for example there are norms that are specific to science

teaching to which the teachers hold each other accountable when they justify pedagogical decisions. This is illustrated by all of a student's science teachers providing them with sufficient experience of practical work in *each* of biology, chemistry and physics so that it is appropriate for use in their GCSE assessments. Third, there is a well-honed repertoire of ways of reasoning with tools and artefacts. This is evidenced in reasoning with schemes of work: biology, chemistry and physics teachers each provide schemes of work for covering their science subject component of the students' Triple Award Science GCSE. The value of the construct, community of practice, is that it brings together theories of social structure that emphasize institutions, norms and rules and theories of situated experience that emphasize the dynamics of everyday existence and construction of interpersonal events (Wenger, 1998, pp. 12-13).

As this research focuses on teachers of different science subjects functioning within science faculties which are within schools, it is useful to adopt Wenger's (1998) view of a school as being too large to be considered as a single entity or 'community of practice' but rather as *'a constellation of interconnected practices'* (ibid., p. 127). As Cobb *et al.* (2003) note, it provides a useful analytical approach that focuses on the functions of the teachers and delineates the communities of practice whose members contribute to the accomplishment of these functions. Consequently, teachers' practices are characterized as activities that are distributed across a configuration of communities of practice within a school that is viewed as a living organization. This is how I experience schools to function. Senior management team staff and ancillary teaching staff are two different communities of practice within schools. From my own teaching experiences I know that as a member of science staff as one community of practice, *my* practices overlapped with those of senior management staff (my GCSE tier entry decisions had to be discussed with them on an annual basis) and ancillary staff such as laboratory technicians (my GCSE chemistry practical assessments for my students had to be jointly arranged by us). Analyses of this type attend to interconnections between the different communities of practice which are *woven* together (Cole, 1996) and in Cobb *et al.'s* (2003) terms, involve boundary encounters. In short, by adopting Wenger's notion of communities of practice that interconnect, I can appropriately analyse what I hear and see when I engage with teachers in their schools.

Lave's (1988) view of *arena* also allows a Vygotskian sociocultural approach to be applied to the schools and teachers in my study. I use Lave's (ibid.) concept of arena to describe a school, namely as *'a physically, economically, politically, and socially organised space-in-time ... within which activity takes place'* (p. 150). A school / arena is *'not negotiable directly by the individual* [for example, a teacher] ... *It is outside of, yet encompasses the individual, providing a higher order institutional framework within which the setting is constituted'* (p. 151). For example, teachers do not directly influence whether their school is for 11 – 16 or 11-18 year old students. However, an arena is not isolated: it is situated and influenced by practices outside of its boundaries (Bruner, 1996; Wenger, 1998). For example, a school needs to adjust to the requirements of government educational reform as in the introduction of the National Curriculum in 1989.

The constructs of community of practice and arena enable me to research teachers in schools as I *understand* them to exist from my own teaching experiences. For example, science faculties within schools each consist of biology, chemistry and physics departments. By using the construct each department can be viewed as a community of practice within the arena of the science faculty, which is itself nested within the overarching, larger arena of the school. Such a view allows for me to analyse for common and different *interconnected* practices. For example, from my own teaching experiences I anticipate that the science subject departments of the schools where I engage with teachers in this study prepare their students in sets for specific tiers of KS3 science SATs. It also allows for discontinuities in practice, for example the physics teachers in a school may teach its students in sets aiming for particular tiers from Year 7 entry whilst the chemistry and biology teachers might delay such setting until Year 8. These practices might reveal the differences between the participants' interpretations of their goals in the context of the overarching goal of preparing students for the KS3 science SATs. The overarching arena, the school, provides the higher-order institutional framework that is not directly negotiable by any of the science subject teachers, for example it was a non-negotiable requirement that *all* of these teachers enter their students for the KS3 science SATs in Year 9[1].

[1] This requirement remains compulsory in England in 2008 but not in Wales.

Wenger characterizes a community of practice as a mid-level unit that does not go as far as the detailed choreography of interactions. However, I need to go this far as I engage with teachers to access their practices and beliefs. For that reason Lave's notion of a *setting* is useful to me. Lave's (1988) definition of a setting as *'a relation between acting persons* [primarily teachers in this research] *and the arenas* [schools] *in which they act'* (1988) is enriched by Nespor's (1997) notion of *'intersections in social space, knots in a web of practices that stretch into complex systems beginning and ending outside of school'* (pxiii). In combination they provide me with a view of a setting within a school as being personally experienced and foregrounding subjective experience; it is orchestrated by teachers but participants and influences from both within and outside of the school are fundamental to what is created and made available to be experienced. Neither the setting nor the participants' activity exist independently, they only exist in relation to each other. Students and teachers bring beliefs and knowledge to settings as a consequence of their participation in a multitude of other social contexts (Murphy and Ivinson, 2004) so that teachers' practice and students' actions are constantly negotiated and evolving: Therefore, in my view, the arena, setting and its participants, which in this research are primarily taken to be teachers, are by their nature mutually constituted. Rogoff (1995) addresses this by referring to three planes of analysis, social community i.e. the arena, interpersonal (Lave's setting) and personal experience of the setting. In describing one plane, it is brought to the foreground, but the other two are always there and have to have attention paid to them. This has major consequences for how I present the analysis and findings from my qualitative investigation - when describing one aspect of this mutually constituted arena, setting and its participants, by its nature I will find I am led to describe others and on trying to describe another aspect, I will find that I am inevitably repeating material from the former.

5.2 Methodological approach: a qualitative case study

If settings are *personal* and teachers' mediation is *personal* as argued by Lave (1988), my primary concern is with teachers' *individual* accounts and mediation. Stake (1994) suggests that actions (here, those of teachers) can only be understood in the context of narrative accounts which draw on the whole culture in which the action occurs so that a cultural perspective can also be understood with other relevant aspects. In case study research a person, an enterprise, an event, an institution,

a programme or a population, a time period can all be considered as cases (Stake, 1978; Patton, 1990). As Patton (1990) suggests, case study is particularly useful for understanding some special people (my chosen teachers) in their unique situation (teachers' schools at the time of my data collection) in great depth. He suggests a great deal can be learned from a few exemplars of the phenomenon in question, which in my investigation is mediation of assessment by teachers. Consequently, I decided to consider teachers as cases studies. Here, *case study* is understood to mean an *'approach to understanding'* (Stenhouse, 1978, p. 24) in which the concern is with the *'situation as a whole'*. In this investigation a case is a teacher within a particular arena i.e. science department or school. The 'situation as a whole' encompasses the arena, setting and its participants and can be found within a teacher's individual account of their practices and beliefs, which are situated and mediated by processes within and beyond the teacher's arena. Each account will contain an intersection of issues of community, social practice, meaning and identity (Wenger, 1998). I interpret each account from my position presented in 5.1 in terms of how the arena *and beyond* constitute meanings and practices. I intend to explore aspects of each teacher's practice for entering their students for science GCSE examinations to understand the reasons for their choice of tier, syllabus and examining group and the influences that might militate against those choices. I wish to explore each science teachers' beliefs about the relative difficulties of the different science subjects, including in relation to the assessment instruments used in the GCSE system, and again, understand how these beliefs constitute their practice in relation to their students and its impact on examination comparability. As noted by Hammersely and Atkinson (1995), although it is epistemologically impossible to give an exhaustive account of any object, a case study methodology still provides *'a means of investigating complex social units consisting of multiple variables of potential importance in understanding the phenomenon* (here, teachers' mediation of assessment). *Anchored in real-life situations, the case study results in a rich and holistic account of the phenomenon'* (Merriam, 1998, p. 41).

To obtain such an account it is necessary to collect relatively detailed data from science teachers. A case study design does not claim any particular method for data collection (Stake, 1978). However, the underlying complexities of the social world of teachers are better explored by a qualitative case study that is descriptive and interpretative in its overall intent. It is 'descriptive'

in the sense that my write up will describe and analyse situations to offer a rich portrayal of the phenomenon under investigation, teachers' mediation of assessment, attesting to the complexities of the situation. The resulting rich, thick description will form the backbone of my 'interpretative' case study. In this process I aim to get close to my teachers within their natural settings. In that sense there is a semblance of ethnography, but as I am unable to spend time in the field observing teachers directly, a case study approach in which I can engage individual science teachers (cases) in discussion (interview) seems appropriate for this purpose. Interviewing is a two-way process and allows me to interact with the teachers, keeping the total context in consideration, remaining open to new insights and being sensitive to data (Merriam, 1998). This enables me to understand each feature of a case in the context of its other features (Hammersley, 1989), helps to keep the investigation 'open', as in the enthnographic tradition, to elements that cannot be codified at the time of the investigation (Baszanger and Dodier, 1997) and facilitates a more probing investigation than could be undertaken with a questionnaire. I considered this advantage outweighed the time that it would take for me to complete the interviews and analyse an expected large amount of generated information. My intention was to *'investigate a few cases ... in considerable depth'* (Gomm *et al.*, 2000). As Shulman notes:

> *To claim that one is conducting a case study requires that an answer be provided to the question, " What is this a case of?" Not every description is a case study. It may be a description of a singular individual or event. To claim that something is a case study is to assert that it is a member of a family of individuals or events of which it is in some sense representative.*
>
> (Shulman, 1981, cited in Wilson and Gudmundsdottir, 1987, p. 44)

In the context of this research, the cases are cases of individual science teachers. There was no intention of defining further, at the outset, what these might be 'cases of'.

My primary intention was to look for the *personal* in each of my cases. However, what appears in the teachers' accounts is mediated by the arena. 'Meanings' are developed together in the teachers' 'communities of practice' (Wenger, 1998) and what individuals appropriate is their understanding of this shared meaning. As all of the arenas are science departments and schools, I anticipate finding 'enduring practices' (ibid.) *within* the arenas because of the subject (science) and

across arenas because they are part of the same type of overarching arena, a school. So I anticipate these enduring practices to emerge in the different teachers' accounts of their practices and beliefs. For example, the influence of cultural and community beliefs about learners and views of mind are possibly shared as appropriated understanding, these having emerged from professional training communities within which the science teachers develop. Adopting a Lave and Wenger (1991) perspective I look first to understand each teacher's account - its 'uniqueness', before looking for enduring practices and shared beliefs. I intend to consider whether there are similarities across the cases and to understand them. I also consider whether there are differences and to understand them too. My theorising is about teachers' practices and beliefs being aspects of an *integral* system in which the nature of teachers' mediation of the assessment process might impact on examination comparability.

Thus my cases are not 'instances of type' (Gomm, *et al.* 2000, p. 4) described in terms of a particular theoretical perspective and my case study findings are not assumed to be generalisations. As noted by Guba and Lincoln (1981, 1982) since social phenomena are neither time nor context free, generalisations are impossible. Their view of the aim of case study research is to produce *'an ideographic body of knowledge'* (Guba and Lincoln, 1982, p. 238), which is best encapsulated in a series of *'working hypotheses that describe the individual case'* (ibid.). They go on to suggest that some transferability of these hypotheses may be possible *'depending on the degree of temporal and contextual similarity'* (ibid.). Based on the idea of *transferability*, they also call for replacing the notion of generalisability with that of *fittingness*: the degree to which the case studied matches other cases. I have adopted this position - that it will be for others to decide whether my case study findings are applicable to other cases than those I have researched (ibid.).

School structures such as the accommodated age range, school processes for allocating students to teaching groups such as banding and setting, and the socioeconomic profile of a school's catchment area are but a few ways by which schools may develop differently to give different meanings to learning and assessment. Different schools can be seen as different social arenas in which teachers and students are positioned in different ways so that assessment practices and teachers' mediation of them can potentially vary significantly. It is for this reason that I chose several schools in which to conduct my case studies. I chose my schools on the basis that by taking

teachers from different schools I envisioned that different structures and processes would be encapsulated in my teachers' personal accounts and in this way I would capture valid glimpses of reality – in short, to better reveal any variety in the ways by which teachers mediate assessment.

5.3 Ethical considerations

As 'the right to know' can easily clash with the principle of respect (Pring, 1984), it was essential for each teacher to give his or her informed consent. My choice of schools was predicated by my knowing professionally and personally several head teachers and science teachers in the locality. Thus I made initial contact with science teachers by telephone to explain my purposes and strategy and to gauge their willingness to be interviewed. I made clear that their interviews would be audio-recorded with their participation made as easy and pleasant as possible. I also explained they would be required to check their interview transcript for its representativeness of their views and practice, and assured them anonymity and confidentiality for all aspects of the interview data. I then sought the consent of the head teachers of the schools in which I hoped to interview my teachers – again this was done by telephone as I knew these head teachers personally. My purposes for interviewing teachers were explained and the anonymity of school and teacher in any related reports was assured. I emphasised my intention to cause the least disruption to the on-going life of the school and made clear that I would not be involving any students in my investigation. I offered the head teachers an opportunity to read any subsequent publications based on my interview data but in all instances this was declined. As each school's interviews were completed, I sent a letter of thanks to each science teacher and head teacher.

Throughout the interview period I checked that I was following my intended strategy. Some of the interviews were transcribed by a third party. I checked the accuracy of *all* of the transcripts and annotated them for emphasis of particular sections of text to better reflect the emphasis given to them by the interviewees. As another accuracy check, I asked my husband to read a sample of my transcripts against the interview recordings. All of the teachers returned their transcripts endorsed as representing their practices and beliefs. To avoid bias in the interpretation of my interview data, I re-read my findings against the interview transcripts several times to check for possible alternative interpretations.

Exploitation occurs when participants in research get little or nothing in return for supplying the researcher with information (Hammersley and Atkinson, 1995). My teachers had my letter of thanks but in trying to give something back to my participating teachers, I acted upon the pay back strategy of Ely *et al.* (1991). This involved me telling the stories of my teachers rather than imposing my own – reporting *their* meanings, and describing *their* social context not as separate but as it was lived and understood by them.

5.4 Research design

5.4.1 Sampling

Given the resources available to me I decided to have a total of nine cases (science teachers). One biology, one chemistry and one physics teacher formed my cases within each of three schools. This enabled me to explore associations between the different science subjects and the teachers' practices and beliefs within the same arena *and* across arenas.

Within each school I decided that one of the three science teachers should be the head of science faculty who, in my experience, knows their whole school policies and practices. This is to allow information about each arena to emerge, for example the policy and practices for students' allocation to teaching groups to aid my interpretation of the science teachers' accounts.

5.4.2 Face-to-face interviewing with a semi-structured approach

I decided to use a semi-structured interview approach for several reasons. It would give me the opportunity *'to probe deeply, to uncover new clues, to open up new dimensions to the problem and to secure vivid, accurate, inclusive accounts'* (Burgess, 1982, p. 107). I had already identified the main issue that I wished to explore with each case, namely, the basis of the judgements made by teachers when entering their students for GCSE science examinations, as a means of exploring their classroom practices and beliefs being situated and mediated by processes beyond school. I wished to retain the freedom to allow teachers to express their views and feelings as fully and spontaneously as they wished. I wished to retain the flexibility to respond to what teachers might tell me during the interview - probing for more depth when appropriate and being able to pursue unexpected answers and issues when I thought it relevant to do so. The issues I wished to investigate are relatively under-researched and I had no associated preconceived understanding against which to elicit the teachers' responses. My intention was to use a number of open-ended

questions related to my technical findings as probes to provide opportunities for teachers to tell me their 'stories' in relation to their practices and beliefs when entering their students for GCSE science examinations. In turn, the teachers' 'stories' would be interpreted for illuminating aspects of examination comparability.

I did not provide any personal opinions on the issues covered during the interviews and this did not diminish my rapport with my interviewed teachers, most likely because I knew all of my teachers socially, though not very well. I was deliberately respectful, non-judgemental and non-threatening. For these reasons a climate of trust was established for the interviews.

5.4.3 Choice of questioning style: an emphasis on open questions of an indirect nature

I chose to ask open-ended questions mainly for the reason identified by Kerlinger (1970) as '[they] *supply a frame of reference for respondents' answers, but put a minimum of restraint on the answers and their expression'*. They also have the advantages of encouraging co-operation and establishing rapport, allowing an interviewer to make a truer assessment of what a respondent really believes (Cohen and Manion, 1991). I also decided to frame my questions in an indirect form whenever possible. As Tuckman (1972) notes, a cluster of problems surround the person being interviewed and I needed to consider the extent to which my questions might influence the teachers to show themselves and / or their schools in a good light, and anticipate and respond with what they think I want to hear. Indirect questioning would reduce the pressure on the teachers to produce rationales for their practice and thinking - rationales that might not be truly held and would reduce the validity of the obtained information. Thus rather than simply asking my chosen teachers what judgements they made when entering their students for GCSE science examinations, I asked several questions of a more indirect nature. I initially focused on asking the teachers what they *do* i.e. their practices, rather than starting with what they believe. These questions trigger / prompt the teachers to articulate their views of, and personal responses to, the practices instigated at an arena level. There are influences that have emerged from the literature about examination comparability that I wished to explore – whether the teachers recognise these and what their position is in relation to them as this, like other beliefs, might be given significance by teachers in their accounts of their practice. Other beliefs that are significant to the teachers I allowed to emerge.

5.4.4 Pilot and subsequent amendments

I piloted my interview questions with a female chemistry teacher from an 11-16 school. She had more than thirty years teaching experience as Head of Chemistry and was the school's Co-ordinator of GCSE examination entries. I chose her for the pilot knowing that: she has considerable experience of entering her students for GCSE science examinations; an understanding of the policies within her school due to her senior management role; could articulate her practices and beliefs, and was willing to give generously of her time. The interview was tape-recorded. I used an aide memoire of the questions and prompts I was to ask as support for keeping the interview on task and to a time limit of one hour. This interview produced my desired outcomes in terms of richness of response to all questions. However, I overran my one-hour time limit and felt that I only completed all questions because I could draw on the friendly relationship I had with this particular teacher. We were both becoming tired part way through the penultimate question. I could not assume I would have this degree of co-operation with my case study teachers. Consequently, I reduced my questions. I omitted a question that asked teachers to comment on the paper construction analysis outcomes (Benson, 1995). This question was seen as a means of eliciting the teachers' perceptions of the cognitive demands and differences therein for biology, chemistry and physics. I chose to omit this particular question because the previous questions had already provided some insights of the teacher's thinking in relation to this issue. On the re-written aide memoire (Appendix 7) I also inserted some prompts for me to pace the time available for each question's response, and reflect on the information provided by the teacher in terms of its research-relevance at different times in the interview to keep me on task and not miss unexpected useful sources of information.

Each taped interview commenced with my re-iteration of the purpose of the interview, namely to explore their beliefs about GCSE science subjects and their practice in relation to their preparation and entry of their students for these examinations. The interview ended with my providing the teacher with an opportunity to reflect on their responses to my questions and to amend / provide additional information if they so wished.

5.4.5 The interviews

Appendix 7 includes the main interview questions and prompts. The first question asked the teacher about the Year 9 classes that they taught. This initial conversational approach was aimed at relaxing the teacher to make them feel more comfortable with the more probing questions that would follow. The purpose of my first question was to identify the teacher's teaching commitments across the different Year groups, the different science subjects and the principles and processes used for allocating students to teaching groups in each Year.

My second question was to explore more deeply the allocation of KS4 students to teaching groups and the teacher's decision making. I planned to encourage the teachers to reveal their practice and associated views. My third question asked the teachers about the GCSE examining group they used and the history relating to this. My interest was in the school's practice and to gain insights of the teacher's interaction with this. Again I used prompts (see* in Appendix 7) to obtain further insights. My fourth question sought to illuminate the teacher's practice and beliefs in relation to their tier entry choices for their students, a primary focus of my research. My approach was again indirect. I referred to the teacher's mark / class book and asked the teacher to tell me how any student came to sit the science examination that they had that summer or were allocated to sit the following summer. I was interested in what aspects of the school, the subject, the students and their parents they considered significant.

My fifth question was more indirect than the previous four. It sought teachers' beliefs about subject difficulty. I presented the teachers with the notion that some people believe the separate science GCSEs are not equally difficult for students and asked them for their view on this. I used three prompts to challenge their view to further explore the teacher's beliefs (see* in Appendix 7).

5.4.6 Data sampling and processing

Table 5.1 identifies the teachers by specialist subject and school. Each teacher is allocated a Christian name beginning with the first letter of their specialist subject so as to enhance their 'voice' in the reading of my chapters.

Table 5.1 The Interviewed Teachers

School	Subject Taught		
	Biology	**Chemistry**	**Physics**
1	Barry	Cathy	Paul
2	Betty	Clive [HOScD]	Peter
3	Brian [CHOScD]	Clare	Phil [CHOScD]

[HOSD] = Head of Science Department

[CHOSD] = Co-Head of Science Department

Each taped interview was transcribed and additional indicators of the teachers' associated actions and attitudes were included. For example, I inserted a star (*) when the teachers' responses to my questions – and any other issues raised by the teachers themselves, appeared to be expressed with a strong conviction. The indicators were to help my interpretation of the interview data and to extend the representation of the teacher's 'voice'. The teachers were each given the opportunity to endorse their transcription as a record of their interview. All transcriptions were endorsed.

5.4.7 Analysis and presentation

For each teacher from the same school I listened to their taped interview and read the associated transcription several times to establish a sense of the whole of the interview. As I did so it was clear that the teachers' transcriptions were reporting how contextual issues interplay with practice at an individual teacher's level. From a situated view a setting is a personal response to the same arena. Therefore, the general school specific issues for each school, the 'enduring practices' of the arena, needed to be characterised first to enable the significance of the teachers' personal responses to emerge clearly. I started the school level analysis using NUD*IST. I discontinued this for a number of reasons. First, a great deal of time was needed for entering my extensive interview data. Second, all categories had to be identified prior to data entry for the purposes of coding and creating nodes and I found this disallowed a situated personal perspective. Within NUD*IST electronic formats I also found it difficult to obtain overviews of sets of interview data. I reverted to using mechanical methods for sorting and ordering my transcript data and adopted Hycner's (1985) 'bracketing and phenomenological reduction'.

Identifying the school specific issues

As Wolcott (1990, p. 33) advises, I began sorting my interview data by *'finding a few categories sufficiently comprehensive to allow* [me] *to sort all of my data'*. My first intention was to use the content of each school's transcripts as the resource to identify school practices. The themes of my interview questions and prompts largely served as a skeleton for the identification of these school practices (Wolcott's *'categories'*). For example I anticipated one school practice would be 'examining group choice', the theme of my third interview question. From reading all three of the school's teachers' transcripts I identified the nature of each school's policy on this issue, for example the identity of the group used by the school, how long it had been used, who had control of its choice and the rationale for its choice. As expected, several practices emerged from each school's set of transcripts as in the case of the concern with the effectiveness of the mathematics teaching in the school, which emerged for School 1.

Identifying the teachers' perspectives.

I explored the transcripts to obtain each teacher's perspective of his or her practice. I looked for connections between practices and beliefs both in the arena of the school and the setting of the teacher. I foreground the intrapersonal within the setting in which the teacher works for each teacher's perspective of his or her practice relating to students' GCSE science examination entries.

The findings are summarised in Chapter 6. The teachers' perspectives are presented in the order of School 1, 2 and 3 and after an introductory section describing the school / arena practices. These practices are organised in relation to Year groups and examining group. In describing School 1 as an arena, issues generic to all three schools are first explained and defined making this a longer section than the others. For *each* school, the teachers' perspectives are presented in order of their comprehensiveness of the number and types of issues raised. This means that many of the issues raised by the remaining teachers in that school have already been described, thus minimising unnecessary repetition and facilitating clear cross-referencing between the teachers' perspectives and responses.

The amount of data available from each teacher varies. This is due to several reasons. Some teachers responded less well to certain interview questions because in their school they had less involvement with the issues. This is particularly true for Betty, the biology teacher in School

2, who had only been in School 2 a year at the time of the interview. She was also the least experienced of all the teachers, having taught for four years. She appeared comfortable with me interviewing her so I do not think that the smaller amount of data I obtained from her was due to me a researcher. Cathy, the chemistry teacher in School 1 was willing to answer my questions but was only prepared to spend a limited time with me. I had arranged to interview the Head of Science Department who is also in charge of Chemistry at School 1 but at short notice, he decided his chemistry colleague should be interviewed instead. Cathy made it clear that she was not prepared to give more than half an hour to the interview and this disallowed full engagement with the questions. This could be due to her not really wishing to give up her non-teaching time and, or, not being fully aware of her school's policies as seemed to be the case when asked about related issues. For these reasons Cathy's report is 'thin' in comparison to those of Paul and Barry in the same school. Clare, the chemistry teacher in School 3 fully engaged with my questions, appeared knowledgeable about her school's policies and provided rich insights into her practice and beliefs. However, Brian and Phil, the co-Heads of Science Faculty in School 3, had respectively particularly strong interests in students' access to learning and assessment processes and provided more detailed responses to my questions than any of the other teachers in all three schools. It is for this reason that Clare's personal response comes after those for Brian and Phil. One could argue that I have been less effective as a researcher with female than male teachers. I do not think this is necessarily true because of the reasons outlined above for Betty and Cathy and I did obtained rich insights of the practices and perspectives of my female pilot interviewee and Clare in School 3.

In the next stage of the analysis I looked *across* all nine of the teachers' personal responses to identify unique concerns, commonalities and their significance. I used the method of constant comparison to look for patterns, for example when distinguishing similarities and differences between teachers, between science subjects and between schools (arenas) (Chapter 7). Finally, I explored the findings from my analysis for evidence of any emerging general issues regarding teachers as orchestrators of assessment (Chapter 8).

CHAPTER 6

Arena and individual mediation of the assessment process: teachers' accounts

6.1 School / Arena 1

School 1 is an 11-16 comprehensive, co-educational school formed from the merging of grammar and secondary modern schools some twenty years previous, in a town that has been economically deprived since the closure of its local mines. A significant number of teachers taught at the school when it was a grammar school with a tradition of entering high achieving students for GCE 'O' level a year earlier than normal. The banding, setting and grouping of students throughout Years 7 – 11 is complex and is described rather than summarised in a table.

Year 7 Grouping Rationale

Key Stage (KS) 2 SATs are the national tests taken by all students in the core subjects (English, mathematics and science) in the May of Year 6 (age 10-11 years old). The KS2 SATs' results for these core subjects are used to produce an average score for each student who are then ranked and allocated to 'mixed ability' registration groups in that each group contains students with a variety of average SAT scores. However, for identifying teaching groups for subjects, the rank order is used to allocate students to three bands. The 'top' band is composed of students with the highest average SAT scores and the 'lowest' band, designated as 'remedial', is composed of students with the lowest rank order average SAT scores.

The Head of Science Department uses the KS2 Science SAT results to further rank order students to allocate *each band* of students to science teaching groups. This is to ensure that all the teaching groups within any band contain students with a range of results for KS2 Science. The top and middle bands contain respectively five and three teaching groups, and the whole of the remedial band is allocated to one teaching group. All students in any band are timetabled for their science lessons at the same time, and each teaching group is arranged to be a *'manageable'* size.

A student's science group placement is therefore dependent on their band allocation based on their average KS 2 SAT result. Furthermore, decisions made at a whole school level about the appropriate numbers of students for each band in any particular year dictate where band divisions are made. For a minority of students whose average SAT outcomes place them close to these

divisions it is not unreasonable to assume that their band allocations are arbitrary, being based more on what is deemed *'manageable'* than on what SAT results might indicate about their potential. All Year 7 students are taught the same science curriculum consisting of separate lessons of biology, chemistry and physics from the commercial scheme, 'Science Now'. The top band is traditionally made larger than the middle band so that within staffing, accommodation and curriculum constraints, '*the maximum possible number of students access to the learning opportunities associated with the top band for all subjects*'. Science teaching groups in the top band are also consistently larger in student numbers than those in the middle band. The School's rank ordering practice results in more girls than boys being placed in the upper band and more boys than girls in the middle band for Year 7. One might speculate that this is due to the girls' better performance in English SATs having a weighting effect.

The Science Department's policy is to use tests from the 'Science Now' scheme for assessing upper and middle band Year 7 students '*because the content of these tests is referenced to SAT tiers and levels'*. This enables the Department to report students' progress in accordance with whole school policy with reference to SAT levels, and to have a predictive SAT level for each student for tier entry decisions in the Key Stage 3 Science SATs in Year 9. The 'Science Now' Year 7 tests have a common core of items and they are also differentiated using other items to form tiers of test papers that are equivalent to specific SAT levels. Top band Year 7 students do a tier of these tests giving them access to SAT levels 5 and 4, whilst middle band students only do a tier of tests which gives them access to SAT levels 4 and 3. The science teachers write tests for assessing students in the remedial band, which has just one group of students. These tests do not necessarily always reference to SAT levels. For the majority of Year 7 students, a preoccupation with SAT levels and associated levels of performance dominate School 1's assessment practice and views of students' progress in science.

Student movements between Year 7 upper and middle bands occur once, half way through the school year. The movements are based on the rank ordering of the 'Science Now' test results, and the test results from the other core subjects, mathematics and English. It is not possible to move a student to a different band for just one subject. As in the case of movements between bands, timetabling constrains students' movements between science teaching groups within bands.

No record of the extent of student movement between and within bands in Year 7 was made available to me.

Year 8 and Year 9 Grouping Rationale

Comments on the Year 8 and 9 groupings are from Paul. In Year 8 students may opt to take a second foreign language (German). The students opting to do so form one band (German band) and the remainder form another two bands (non-German bands), the smaller of which is designated as 'remedial'. The German band and the larger of the non-German band of students are each divided into science teaching groups based on their 'Science Now' Year 7 common core test results. These science teaching groups are referred to as 'sets' because they are differentiated – for each band, students with the highest test results are allocated to the 'top' set, and so on down to the 'bottom' set containing students who performed least well in the tests. In practice students in the 'top' set of the German band and those in the 'top' set of the non-German band are not comparable in terms of their attained science SAT level profile. There is a preponderance of girls in the top science sets in both the German and non-German bands.

The KS 3 SAT is the national science test taken by students in May of Year 9. The sets within each Year 8 band are associated with a particular tier of the national SAT science papers. All sets follow the same course but pitched at different levels, for example at SAT level 6 for middle sets and SAT level 7 and extension for top sets. Movement of students between the German and non-German bands in Year 8 rarely occurs and when it does, it is largely for social reasons. Movement of students between sets within a band in Year 8 occurs once and usually at a time no later than half way through the school year. Such movements occur to enable all students within a particular set to be aiming for the same tier of the national KS3 SAT papers. The number of students in the top sets in both bands is kept deliberately high to enable as many students as possible to be entered for the higher tier of SAT paper. Nevertheless, student numbers is an issue for class management and whenever a student is moved up into a higher set, usually a student is also moved down from that set. The same banding and science setting systems operate in Year 9 as in Year 8. The vast majority of students remain in the same bands and science teaching sets throughout Years 8 and Year 9.

Year 9 students sit a mock SAT to confirm their KS3 SAT tier entry. As a result *'a few'* students do not sit the same tier as the rest of their set members. Records show *'that more of the sets in the German band than the non-German band are entered for the higher tier of KS3 SAT science test paper'* and that over time they consistently achieve higher KS3 SAT results than the non-German band students. The allocation of a student to a science set and to a KS3 SAT tier is based on their average performance across science subjects. Again no account is taken of differential performance in biology, chemistry and physics.

Year 10 GCSE Grouping Rationale

A student's performance in the separate science disciplines only becomes an issue for group allocation when GCSE science course decisions need to be made in Year 9 for Year 10. Students commit themselves to their Year 10 courses and timetabling arrangements are in place before the KS3 Science SAT results are available. The School analyses the KS3 SAT science results into their biology, chemistry and physics components for each student. These are used both to inform and justify teachers' allocation of students to Year 10 science courses. For example, parents wishing their child to take Triple Award GCSE Science are given the KS3 SAT physics score to advise them to enter their child for Double Award GCSE science when they are unlikely to achieve at least a grade C on the Triple Award GCSE physics examinations. The rationale for this advice is a poor performance in any one of the science disciplines is subsumed within an average score from *all three science* disciplines in Double Award GCSE.

Students rarely follow courses other than those advised by their science teachers. Rare contrary instances occur when parent's wishes rather than student's wishes go against this advice. Then the school agrees to enter the student for the parent's chosen course but continues to monitor the student's progress and offer advice to the parents about what is best for their child. The science KS3 SAT analysis outcomes are used to guide students who may be undecided about taking up their chosen science courses at the start of Year 10. Notably, at the end of Year 9 proportionally fewer girls than boys in the top sets of the Year 8 bands opt for the Triple Award GCSE Science course. Usually all students in the remedial band of Year 9 only follow courses leading to the Certificate of Education (CoEA) in Years 10 and 11.

The science KS3 SAT analysis outcomes are also routinely used to allocate students to the Triple Award teaching groups. A high SAT level in all three of the science disciplines secures entry to the Triple Award GCSE course. Each subject in the Triple Award course has two sets and students are allocated to these on the basis of their subject specific SAT score. In contrast, it is the rank order of the students' average SAT science result that is used to allocate them to one of three sets following the GCSE Double Award course.

The set 1^s in the Triple and Double Award courses are made larger than the set 2^s to accommodate students who are judged to be 'borderline'. To this extent, teachers have developed a practice to alleviate the constraints imposed by timetabling, accommodation and staffing resources on students' potential achievements. No student movement between Triple and Double award GCSE science courses occurs in Years 10 or 11. The rationale is that these courses run at different times in the school timetable and such movement would necessitate changes in the students' remaining subjects.

KS3 Science SAT results determine students' science learning opportunities in KS4 at School 1. This is even more significant when one takes account of the neighbouring tertiary college's view reported as *'Triple Award science students are better equipped than their Double Award counterparts for progressing well on 'A' level science courses'* (Cathy). In one particular year, additional time was allocated for the teaching of science subjects in Year 10. It was decided that all of these Year 10 students opting for science subjects would take Triple Award GCSE science because the students were regarded as particularly *'able'* by their science teachers and the increased amount of teaching time enabled them to give additional support to students whenever it was required. This further substantiates the view that decisions regarding students' allocation to science courses and therefore their potential achievements in science are significantly influenced by timetable constraints and opportunities.

The relative numbers of girls and boys in the Triple Award biology, chemistry and physics teaching sets in Years 10 and 11 varies over the years. This variation has been noted for Physics and Biology but not for Chemistry. Contrary to national trends (Institute of Physics, 2006), approximately the same number of Year 9 boys as girls opt for the Triple Award GCSE Science course. However, for physics there are proportionally slightly more girls than boys in the top sets

and more boys than girls in the bottom sets in both Years 10 and 11. In biology twice as many girls as boys were in set 1 in Year 10 at the time of the interview. Thus the disproportionately greater number of girls as boys in the Year 7 top band and in the Year 8 and 9 'top' science sets is not sustained to the same degree in the top sets for Years 10 and 11 Triple Award GCSE Science subjects.

Each Triple and Double Award set prepares for a particular tier of the GCSE examination. The choice of tier is established for Triple Award for physics and biology sets in Year 10 but not until the end of Year 10 for chemistry when the majority of work common to all tiers has been covered. All Triple Award students are entered for GCSE Single Award Science at the end of Year 10. This practice reflects a previous tradition of entering 'able' students for 'O' level GCE a year earlier than usual. Each student's tier of entry for this examination is decided by a consensus of their biology, chemistry and physics teachers' opinions. The School then analyses the Single Award GCSE Science results by performance on the biology, chemistry and physics components. The science teachers use these component scores to make decisions regarding the tier of entry for the student's Triple Award GCSE Science examinations. Some movement of students between the sets for each science subject occurs. Consequently, students may be entered for different tiers of papers for the three different science subjects. In this respect these arrangements facilitate access to the highest possible GCSE grade in each science subject – that is, as judged by the students' teachers. This is in contrast to the policy described above for KS3 Science SAT tier entry decisions, when teachers use a student's average score for the science disciplines.

Examining Group Choice

The School uses (2002) Northern Examinations and Assessment Board (NEAB) for its GCSE science syllabuses and examination entries. The locus of choice of GCSE examining group is entirely with the science teachers who reach a consensus view. The School moved from WJEC to NEAB some ten years ago, a move largely prompted by the science teachers' shared desire to move to a modular course and WJEC not having its support materials available for the then recently introduced national programme of assessment of students' practical coursework. There were some shared concerns regarding the outcomes of NEAB's moderation of students' practical coursework

as teachers consider that too many marks are deducted from their assessment of students. This issue is under review pending making a formal complaint to NEAB.

6.1.1 The Physics Teacher's Perspective and Personal Response: Paul

Paul is in his late thirties and a physics graduate. He is responsible for all physics teaching within the school and has been at the School for three years.

Paul is concerned that to some extent, students with similar abilities may find themselves either in the top or middle band of Year 7 – allocation appears arbitrary to him. He views the School's allocation of students to Year 7 bands as causing problems for his teaching and assessment of physics. In his experience this allocation leads to the top band of Year 7 students having *'tremendous variety* [in ability]'. This causes him problems in differentiating work for teaching groups and in particular for him responding to students' varied literacy skills. Paul deals with this by writing booklets that differentiate the work and its literacy requirements for each band and the teaching groups within them. Paul expressed concerns about the infrequency of student movements between Year 7 bands. In his view '*we should move them* [Year 7 students] *twice ideally*', rather than the current once, to keep abreast of students' progress. He acknowledges that he has no influence on current school practice in this respect.

In Paul's view the 'Science Now' tests are useful because they are referenced to the national Key Stage levels and correspond with school policy to report students' achievements by levels. He views these tests as being too demanding for the lower band of students designated as 'remedial' in Years 7 to 9. Paul sees this as '*a problem, in that it is difficult to give levels to those tests because we have made them ourselves'*. He uses '*other materials to give an idea* [of levels]' for these tests. This desire to allocate levels to the tests is because '*there is difficulty of comparison between SEN* (statement of educational needs students) *groups and main groups because if you have to move students between them, there has to be some kind of a comparative measure'*. Comparability in assessment is a key concern for Paul.

Paul has responsibility for allocating Year 8 students to their science sets and identifies *comparability* of assessment information as a problem in this allocation. This is because '*it is difficult to make a comparison between upper and middle band students because the upper band do a slightly different testing arrangement to the middle band*'. He uses only the common core results

of these tests, together with students' *'overall course mark, which is comparable, and the end of Year 7 exam marks'* to produce a rank order as a basis for students' allocation to Year 8 sets. Paul also views timetabling and class management as constraints on allocating students to sets. He commented on the arbitrary delineation of students to keep classes to a manageable size. He attempts to mediate this arrangement by considering other factors that might limit students' scores, for example students' underachievement due to dyslexia and recent entry to the school. The decisions are made collaboratively – *'there has always to be a little flexibility and that has to be professional judgement really and I would get together with the rest of the* [science] *staff'*.

Paul thinks it is important to allow students to move between sets, although he recognises the importance of maintaining continuity of teachers too. In his view more movement of students between sets in Year 8 than in Year 9 is acceptable '*because of the time* ... [the] *Year 9 course is less than two terms really and so you have less chance and less testing opportunities to actually fall back on and make a decision about movement*'. He believes '*a group* [of students] *work at a certain level and ... that students who are not so able perhaps, pull themselves up*' to conform to the level of the set. He views placing students in the *'right'* set as '*very important actually ... especially as you might well end up entering the bulk of a set for a paper* [Tier of SAT levels] ... *and students might be mis-entered*'. This indicates Paul's awareness of the potential source of invalidity associated with differentiated entry. Paul feels differentiation in teaching is difficult within any Year 8 or Year 9 set and that generally lessons are pitched at a particular level, level 6 for a middle set and level 7 for a top set. His response is to make top sets in Years 8 and 9 deliberately larger than normal to give as many students as possible access to higher level SAT papers.

Paul does not offer any comments in support of or against his school's policy of entering all Triple Award students for Single Award in Year 10. He does not routinely use the outcomes in his decisions relating to tier entry for Physics GCSE. There are two tiers, foundation and higher. He bases his decisions on '*the coursework ... and all the tests they* [students] *have done* [in physics] *in Year 10* ... [and] *in Year 11*. He gives the students '*a grade according to how they did at foundation and at higher and* ... [sees] *how they are operating in each* [to] *come up with a grade overall had they sat either tier and whichever grade is best'* determines the tier entry. He

tends to choose the foundation tier when there is no advantage in the two grades. He only uses the Single Award Science results to respond to any disparity in students' biology, chemistry and physics tier entry decisions. According to Paul, such disparity occurs for about five percent of students in any Year 10 cohort. Then the response is again to enter the student for the foundation rather than the higher tier. Paul's experience is that about five per cent of Year 11 parents challenge his tier entry decisions. He responds by providing parents with information from *'the wealth of experience* [he has] *of doing loads of tests where they have had the two tiers built in what they operating best at - also all the homeworks I give them are based on past papers and they are graded according to higher or lower tier test papers'*. He is clear about only providing such information and giving the final decision to the parents.

Paul used the NEAB examining group at his previous school too, which had changed from Nuffield. He justifies his choice of NEAB as his preferred examining group with comments that they are a *'very fair board* [group]' and that their examinations are *'fair as well'*. When prompted to expand on what he means by *'fair'* he draws on his experience of using the Nuffield examining group, which he refers to as '*only suiting a certain type of student who is highly motivated'* and to producing *'mismatches between syllabus and examination style'* so that students are *'confronted with a style of questioning which throws them – not in terms of their ability (or) knowledge but in their ability to comprehend the question'*. He views NEAB as *'fair'* as it avoids these practices. Here Paul is showing a concern for construct irrelevance and it determines his choice of examining group.

Paul feels that it is impossible to make valid comparisons of the difficulty of GCSE Physics papers across time, largely because examination papers have needed to change with the introduction of the National Curriculum and the subsequent amendments to it. Nevertheless, he does not '*think the demand of the papers has become less since the last orders came through of the National Curriculum*' and he '*certainly* [doesn't] *think the demand over the past two or three years has changed at all*'. However, since the mid-1980s and so even before the introduction of GCSE, he believes that '*extended writing seems to be something which is disappearing from Physics* [in national 16+ examinations]. He believes advisors and inspectors are still looking for extended writing in students' work and for this reason he still includes it in his schemes of work, but in his

opinion, it is not examined in GCSE physics. He also views the style of questioning within GCSE physics examination papers as having changed across time, with them now being '*easier to understand*' and '*more structured*' – so that students are '*taken through the question now as opposed to an open-ended question*'. His views appear to concur with my technical findings showing a shift towards a decrease in severity of grading in WJEC and SEG GCSE physics from 1995 and this shift being associated with the majority of the WJEC physics examination papers' marks being weighted to just recall of knowledge rather than application, analysis, and evaluation type demands.

Although initially stating that he does not view physics GCSE papers as reflecting any decrease in demand over time, later he makes several statements to counter this view. He states that the calculation work in physics GCSE examinations is '*slightly less rigorous*' than it used to be in examinations pre-GCSE. He views the questions involving physics equations as being as demanding as ever but the computations as being simpler with '*many questions* [having] *numbers which will compute easily*' and certainly not requiring *'to have answers to decimal places ... as you might have in the past'*, and that *'students are now led through a calculation'*. Furthermore, questions from GCE 'O' level physics papers (pre 1989) are viewed as being more like GCE 'A' level standard than GCSE – not that he views 'A' level as now being much easier but that in his teaching he has *'used 'O' level questions from 10-15 years ago when teaching 'A' level and students have found them quite demanding'*.

Despite his reference to GCSE physics calculation work as being more structured and less computationally demanding than its GCE counterpart, Paul believes his Year 10 and 11 students still find physics to be mathematically challenging. In response and because he recognises the importance of mathematical skills for achieving well in physics assessments, he and his physics colleague offer short mathematics courses for students. My technical findings showed that compared with biology and chemistry the students' WJEC performances in mathematics were most positively correlated with performances in physics. A high performance in mathematics was predictive of a high performance in physics. Both boys and girls voluntarily attend Paul's additional mathematics lessons but proportionally more girls than boys do so. He has also produced a 'Maths Required for Physics' booklet for his students. Off tape Paul stated that there

was a problem with the effectiveness of mathematics teaching within the school and that in his view this was reflected in the disproportionately low number of mathematics GCSE grades A*- C and the number of students receiving private tuition in mathematics. However, in his view, students would still find physics mathematically challenging even if the effectiveness of the mathematics' teaching were improved within the school.

Paul spoke at length about students and their parents having *'baggage'* about physics, that it is seen as a *'man's subject'* because it is to do with *'engineering type things'* and biology is perceived as *'cleaner'* and *'more of a women's subject'*, which reflects research findings on images of occupations and gender in occupational stereotypes (Glick *et al.*, 1995). For this reason, in his view, girls come to physics lessons believing that they are going to find the subject hard largely because *'they perceive physics as being mathematically demanding'*. This view of girls' lacking confidence in mathematics is supported by research (TIMSS, 1999, 2003; Gill, J., 1994). My technical findings showed that girls underperformed boys in physics for all my WJEC populations and for one of the two SEG populations despite outperforming boys in their average GCSE grade scores to a significant level (0.1%) in all WJEC and SEG populations. Boys outperformed girls in GCSE mathematics for all of my WJEC populations, but the tier these students had been entered for was unknown and so reduced the dependability of this finding. My physics: mathematics correlation coefficient values for the two sexes were sufficiently similar as not to offer an explanation as to why the girls achieve significantly less well than the boys in physics. This phenomenon did not appear to be related to the relative performances in GCSE mathematics. Girls' lack of confidence was shown to disadvantage them in physics' APU assessments (Johnson and Murphy, 1986). Arguably from Paul's interview, girls' lack of confidence in physics, and in perhaps *its* mathematical demands rather than mathematics per se, contributed to girls underperforming boys in physics in my technical investigation.

Paul recalls advising at least two to three girls per year that it is appropriate for them to enter the higher tier physics GCSE paper but then they insist on entry to the foundation tier. My populations were selected on the basis of students being entered for higher tier in all three Triple Award science subjects. They all consisted of approximately two thirds boys and one third girls. This could be due in part to more girls than boys being entered for foundation tier; I have no

evidence suggesting that Triple Award science courses consist of two thirds boys and one third girls, but rather from the teachers' interviews that approximately the same number of boys and girls take these courses. Paul attributes girls' reticence for higher tier entry to their lack of confidence in their ability and does not find this with boys. He also observed that the proportionally greater numbers of girls in the top science sets in Years 8 and 9 are not sustained in Years 10 and 11GCSE physics sets. Occasionally there are more girls than boys in the Years 10 and 11 top physics sets but it is more usual to have equal numbers. He offered other views of girls and boys differences in physics lessons: girls present better work; they generally work harder; they produce better course work because *'they are prepared to put the time into it'*. In his view boys seem less phased by *'heavy mathematics and the deeper understanding of science'* and that '*sometimes ... the girls put the shutters up as soon as the subject becomes more demanding and this is something we are really working hard to overcome. Certainly sometimes girls say 'I can't do this Sir, it's all maths.*' He responds by encouraging girls to attend the extra mathematics lessons referred to above and generally providing girls with more verbal support than boys during his teaching.

6.1.2 The Biology Teacher's Perspective and Personal Response: Barry

Barry is in his fifties and a biology graduate. He is responsible for all biology teaching within the School. He has taught at School 1 for 30 years and since it was a well-established grammar school.

Barry says he is not aware of how students are allocated to their Years 7 and 8 bands and Year 7 science sets. He has no input on these decisions and does not see the need to understand them. When describing his own Year 7 teaching groups, he only refers to them in terms of SAT levels, for example groups in the 'top' band as levels 4 and 5, rather than by the band's and group's name. This again illustrates how SAT levels dominate the teachers' thinking.

In his view, students' movements between Year 7 science teaching groups occur more often due to *'Head of Lower School's decisions regarding students' behavioural problems'* in a variety of subjects than for academic reasons. Such movements are not viewed by Barry as *'having any effect on students' progress in science because all groups cover the same units of work, albeit in a differentiated manner to take account of students' learning needs'*. He says he is more concerned about not being involved in the decision making and not being kept informed of such movements– students who have *'not been any trouble at all'* to him are suddenly not appearing for

his lessons. In the way that he talks about his colleagues on these issues, he reveals the school's senior management team's and its science teachers as different and separate 'communities of practice' (Wenger, 1998). He is also concerned about the impact unanticipated student group movements have on his class management. For example, he needs to rethink his organisation of practical work for a changed number of students and consider whom amongst the remaining students might be adversely affected. Students' behaviour and potential to learn are strongly linked for Barry. He views effort as mediating ability and this is another source of *incomparability*; students' motivation to learn is met with opportunities to learn and consequently, access to courses and higher tier examination entry.

Like Paul, Barry views the 'Science Now' tests as useful because they are referenced to the national Key Stage levels but also sees them as too demanding for the lower, remedial band of students in Years 7 to 9. Barry shares Paul's desire to use tests referenced to levels to provide a comparative measure to inform decisions about students' allocation to groups in Years 7 to 9. Comparability in assessment is a key concern for both teachers.

He has delegated responsibility for allocating Year 8 students to Year 9 science teaching sets and in his view *'because Year 8 students are already placed in ability groups for science lessons'*, there are no major set changes from Year 8 to Year 9. Like Paul, he deliberately creates Year 9 'top' sets that are proportionally larger (maximum 33 students) than 'bottom' sets in line with general school practice. His rationale for larger top sets is that there is a finite number of students timetabled at a particular time and *'lower ability students are best managed in smaller groups because they have more behavioural problems'*. Barry uses the same rationale when he allocates Year 10 students to their GCSE biology groups, set 1^s being larger than set 2^s – '*Set 1* [Year 10] *is bigger, yes, yes because they are more able they are usually more amenable and easier to teach so you can have a larger class'*. Both Paul and Barry perceive timetabling and class management as constraints on allocating students to teaching sets and respond by creating larger 'top' than 'bottom' sets showing a concern for access and management. Paul appears more preoccupied than Barry with getting the setting *'right'* so that the bulk of students in a particular set are entered for a tier of KS3 SAT papers commensurate with his beliefs about their ability.

Social influences mediate Barry's decisions about student movements between sets based on rank ordering of assessment marks in the transition from Years 8 to 9. For example, he takes staff and parental wishes into account when their concern about negative peer interactions indicates that movement of students between teaching sets is advisable. He also sees the use of rank ordered assessment marks as being crucial objective evidence for defending decisions about students' set placements when he gets a query '*if someone is promoted and somebody else isn't that thinks that they are as good or better than the person that was promoted and this is why we have to use concrete factors* [the rank order marks]'. In this respect he shares the view and practice of Paul.

Barry talks about students in the remedial band as possibly getting '*a very high mark* [in their tests] *because of the way that we produce their exams and the tests* [multiple choice is used] *where it is easy for them to score marks so you could not really justify one going up because of the high mark'*. He refers to these students as having writing that *'may be very, very poor... and they might get fairly high scores with multiple choice type questions and they might be quick on the uptake and be very good orally but if put into a higher group they would really struggle on the amount of written work that they have to do, so there are those considerations'*. This suggests he sees writing as a source of construct irrelevance, as does Paul. Barry gives less value to students' scores on these multiple choice tests than to their writing skills. He does this for all students, not just those designated as 'remedial'. This practice emphasis is not shared by Paul or Cathy. This may be because biology as a subject - and its GCSE examination papers, has a greater demand for students to write continuous prose than either physics or chemistry. Barry raises no objections to the school policy of routinely only entering Year 9 remedial band students for the Certificate of Education Achievement (CoEA) rather than GCSE Single or Double Award Science in Years 10 and 11. Paul did not comment on this issue.

His rationale for there being no movement of Year 10 students between Double and Triple Award biology courses is that the former is modular and there is a mismatch in the sequencing of the work rather than to differences in course difficulty. On the other hand he encourages movement of students between sets *within* each of the Triple and Double Award biology courses '*as a sort of reward for anyone in set 2 that is doing exceptionally well and to keep set 1 students*

on their toes to avoid being moved down'. He appears to value and use the school policy of allowing set movements within a course for student motivational reasons more than Paul.

Barry prepares each teaching set of Year 10 and 11 students for specific tiers of examination papers as does Paul for physics. Paul and Barry share the view that teaching two tiers of work to the same set of students presents differentiation difficulties – it's *'really awkward ... the difficulty of the work is absolutely vast'* (Barry). Barry refers to students' access to commercially produced revision guides as enabling them to identify what work needs to be known for a particular tier. In his view lower tier students do not then pay attention and disrupt the class when higher tier work is being taught within a set being prepared for both tiers. Thus again, Barry refers to the issue of class management of students' behaviour as influencing his practice.

Barry differs from Paul in how he attempts to make access to higher tier entry possible. Paul sees access achieved by accurate differentiation. From several points made by Barry he views differentiation as relatively crude and difficult in practice. For example, he regards his current set 2 in Year 10 as problematic. He equates a significant number of students in this set as level 5.5 and capable of achieving grade C in biology GCSE. They cannot be accommodated in a set aiming at higher tier (GCSE grades A*-C) due to staffing and timetable constraints. He has not responded to this situation with differentiated work within the set but by teaching this set 2 *'everything, the higher and foundation work'* and will delay deciding individual student's tier entry until January of Year 11 so that *'students have a good chance to prove they can do higher tier work'*. In the interim, no movement of students between groups is envisaged - only if a set 1 student looses motivation and misses lessons resulting in demotion to set 2 and foundation entry. So even in this situation Barry is again using students' set movements to influence their motivation.

Parental pressure to enter students for a higher tier in the Year 11 GCSE examination entries rather than Barry's recommended foundation tier is a '*common*' experience for Barry unlike that for Paul (5% of Year 11 parents). This applies to student's Single Award Science GCSE and Triple Award Biology GCSE tier entry decisions. He regards such pressure as being caused by parents' *'inflated view of their child's ability'*. In these circumstances he relies on his records of students' marks from KS 3 SAT results and coursework (work done in class throughout the school year) equated to GCSE grades to substantiate his decisions. Unlike Paul, Barry uses the outcomes

of Year 10's Single Award Science GCSE as a major indicator of which tier a student should be entered for in Year 11. However, like Paul, he also refers to their KS3 SAT biology component result to validate these tiering decisions. He does not refer to the students' ongoing biology GCSE practical coursework marks as a factor in his tiering decisions – a practice used by Paul.

Barry recalls using WJEC science examinations in School 1 during the mid 1980s. In the late 1980s he recalls observing that WJEC's small size resulted in materials, inset, moderation and approach to AT1[1] not being of such high quality as those of the larger GCSE examining group, NEAB. In his view WJEC's publications always came out later than those of other GCSE groups and so did not give science staff as much time to prepare for teaching scheme changes. According to Barry, NEAB's materials were also more supportive of teaching science at a time of National Curriculum changes and the introduction of GCSE and continue to be so for several reasons: '... *the stuff you get from NEAB was always glossy and a lot of information was given, a lot of expansion of the syllabus ... for AT1 in particular they gave us lots of help and lots of meetings when the new sort of type of coursework was being set up, not the AT1 but the step before that. Let's just say there was lots of feedback and they seemed a very friendly exam board and gave lots of support and good materials. The AT1 scheme in particular was far too complicated with WJEC compared with the NEAB set up and that's why we plumped for the NEAB'*. Barry does not regard it as easier for his students to obtain higher GCSE grades with NEAB than WJEC - this was not a contributory factor in changing GCSE examining group. Like Paul, Barry's rationale for choice of examining group is concerned with 'accessibility', albeit in terms of different issues, for example quality of examining group for Barry and items that do not contain construct irrelevance for Paul.

However, Barry has concerns regarding NEAB's moderation of GCSE science practical work, an issue not emphasised by Paul. The moderator is a chemist and according to Barry, *'he has been very, very strict with the Chemistry work – he has cut their mark. I think last year or the year before he cut them by an average of seven per cent per student, which was a lot. He did not touch the biology course work at all and our perception was that because he was chemist he was being really strict with the chemistry but he was not moderating and not bothering too much* [with the biology and physics]'. Barry refers to collaborating with his teaching colleagues to *circumvent*

[1] AT1 refers to the practical skills component of the National Curriculum and GCSE examinations during the 1980s and early 1990s.

this perceived obstacle by only submitting biology and physics practical coursework for moderation - this being possible within the administrative arrangements for NEAB's Single Award but not the Double Award science GCSE. He recognises this practice as *'putting extra pressure on people* [science teachers] *doing physics and biology ... because they have got a lot more marking to do'*. If he thinks *'they are hit hard this year again'*, which he equates to numbers of students missing high grades such as A*, because of the way in which the moderator deals with their practical work, he will seek change. This might be a request for a new NEAB moderator or a change of examining group. This is another way students' achievement in GCSE Science examinations is mediated by School 1's teachers. Due to the greater emphasis Barry gave this issue during the interview compared to Paul, he clearly feels very committed to this type of mediation.

Commenting on the 'relative difficulty' of the separate sciences at GCSE, Barry places them in the order of chemistry, the easiest, then biology and physics. He perceives this as the pattern in the School and at national level. My research concurs with this view of chemistry as it was shown to be the least severely graded subject in 1993 and 1994 but physics was only the most severely graded in 1993. My investigations showed this pattern changed from 1995. Barry's view is based upon GCSE examination results, not upon students' views of the subjects. He claims this order of 'relative difficulty' has held for the last three to four years within the School. He says biology *'came out on top'*, meaning students generally obtained better grades in biology than chemistry or physics some five to six years ago, but then the biology syllabus became *'a lot more demanding'*.

He sees changes in biology syllabii and examination papers throughout the 1990s as altering the difficulty of the subject for students which created some problems for teachers. He sees progressively more material being pushed into the syllabus for the higher tier and insufficient time to cover the syllabus for all tiers and expressed great concern about not spending *'the proper amount of time'* on AT1s (practical coursework component). He does '*a minimum number of AT1s simply because we* [School1's biology teachers] *would never ever finish the syllabus if we spent the proper amount of time doing it*'. This time / syllabus issue was not raised by Paul for physics.

He also commented on the reduction in *'straightforward revision type questions'*, such as labelling of diagrams, and an increase in data handling type questions in the examinations. My research showed a decrease in the percentage of marks allocated to recall of knowledge from 1995 but only for higher tier biology. My technical investigation showed no noticeable change in severity of grading for biology from 1993 – 1995 but a greater similarity in the weightings of cognitive skills on all three subjects' examinations from 1995.

Barry sees the use of examples in examination questions which require the application of knowledge not referred to within the syllabus, and therefore unfamiliar to students, as increasing the degree of construct irrelevant variation in recent years. He sees this occurring most often in the examples of animals and plants in environmental questions. Barry welcomes the use of examples that are not familiar to students in questions if they are intended to test students' ability to apply their knowledge. He is concerned about teaching biotechnology because of the lack of available biotechnology textbooks and relevant information. This is made more difficult because examination questions are based on the technological details of processes such as fermentation where there are several design alternatives and there is no specification about this on the syllabus. The inclusion of concepts such as DNA coding that were *'never dreamt of'* at GCE 'O' level he views as *'difficult'*, and he sees biotechnology as *'too broad'* a topic and *'demanding'* for students and increasing the cognitive demand of GCSE biology. Conversely, the virtual disappearance of continuous prose type questions on the biology examination papers has made biology easier in his view.

Barry refers to his son's chemistry from another school as evidence of a lot of the more difficult parts of chemistry having been removed from syllabi and examinations in the past three to four years. He specifically refers to the removal of chemical calculations in recent times as a reason for his perception of chemistry being the *'easiest'* of the science subjects at GCSE. This view is not supported by my technical findings which show chemistry having become more severely graded from 1993 – 1995 for WJEC and 1994 – 1995 for SEG. Barry differs from Paul: Barry has strong views about the science subjects changing in severity of grading across recent years while Paul holds the view that it is not valid to make such comparisons, largely because of syllabus changes. Barry views physics as always having been the science subject for which *'it's*

the hardest to obtain a good grade' because *'there's a lot more maths'* and *'a lot of pupils are put off by the calculations'*. In this respect he holds the same view as Paul. Barry also says '*it's more difficult to make physics relevant and interesting than the other two science subjects*'. He believes students find physics difficult and *'not really relevant'* and that *'biology is probably the most relevant because everyone has a body'* then '*chemistry is the next'*, although he doesn't *' really know why but certainly ray diagrams for light and so on – the students wonder why are we doing this and it is really difficult to make it stimulating'*. This is his view of the subject, not just the associated examinations and reveals his view that the science subjects are incomparable. Paul did not raise the issue of relevancy of subjects as an issue when discussing the relative difficulty of the science disciplines. However, he did say that girls have preconceived ideas about physics, that they anticipate they will find it difficult. Interestingly, Barry did not raise any comment about girls or boys and finding any subject easier or more difficult, despite his stereotypical views on the science subjects stated above.

6.1.3 The Chemistry Teacher's Perspective and Personal Response: Cathy

Cathy is in her forties and a chemistry graduate. She does not hold a post of responsibility within the chemistry department and has taught at School 1 for ten years. I had arranged to interview the Head of Science Department who is also in charge of Chemistry but at short notice, he decided his chemistry colleague should be interviewed instead. Cathy was not prepared to give more than half an hour to the interview and was somewhat reluctant to answer the questions. For these reasons Cathy's report is 'thin' in comparison to those of Paul and Barry.

When asked to list her Year 7 teaching groups, Cathy only distinguishes between 'top', 'middle' and 'bottom' banded groups and in terms of their SAT levels, like Paul and Barry - *'the upper* [top] *bands are largely students who have had three 4s* [levels] *in the core subjects and the middle bands with less than that, that's more or less the criteria used* [to allocate them to their band]'. She emphasises the importance of using teachers' consensus view of students' progress in science to place students in sets in Year 8 and 9. She views the current student placement arrangements as not allowing students to be in sets that accord with their 'abilities' in the different science disciplines. She sees timetabling constraints as the reason for the current setting system. However, she considers that relatively few students demonstrate significant differences in their

progress in the different sciences and for this reason views the current setting system as acceptable. Like Paul and Barry, Cathy values the 'Science Now' tests for being referenced to the national Key Stage levels and enabling her to fulfil school policy of reporting students' achievements by levels but she also sees them as too demanding for the lower band, remedial students in Years 7 to 9. All three teachers wish to allocate levels to tests to provide a comparative measure for deciding which students to move between groups in Years 7 to 9. Comparability in assessment is a key concern for all three teachers.

Although Cathy considers all Year 9 students have a free choice over their uptake of Double or Triple Award Science courses, she feels that in common with Paul and Barry, most of her students and parents follow her advice. Unlike Paul and Barry, she emphasises the modular nature of the Double Award course as being more appropriate than Triple Award for *'someone who is better able to learn small amounts of work'*. Barry had emphasised the Double Award's modularity as preventing students' movements between Triple and Double Award courses, whereas Paul had not referred to modularity at all in his interview.

Cathy, like Barry, views sufficiency in teaching time as a key issue in examination entry decisions. She shares Paul's view that increases in teaching time across Years 10 and 11 in the two previous academic years facilitated the entry of proportionally more students for Triple Award Science GCSE and says *'that time has not been generated again'* and that even when it was *'it wasn't really quite enough to cover the three separate sciences without having to race through the work'* with the increased number of students. Thus Paul, Barry and Cathy all share a view of timetabling constraints as limiting their capacity to respond to their students' learning needs and restricting the types of GCSE science courses they are able to offer their students.

Cathy considers the same constraints of timetabling and staffing also prevent movement of students between teaching sets aiming for particular tiers of GCSE science papers. By the beginning of Year 11 she aims to have chemistry students in sets designated for a particular tier of GCSE chemistry papers. She recalls that in Year 11 *'after the mock exams and when the final decisions* [tier entry] *were made ... there was some movement* [between sets] *so that all of the people doing higher tier were in one group* [set] *... because of numbers, it was three students who were doing foundation in that set 1 as well, it just wasn't possible to move those to set 2 because*

set 2 had grown to 34 at that time'. Her response to this was to teach higher tier material to the majority of the students in the top set and give alternative work to the minority of students in this group aiming for foundation tier. This meant that these foundation tier entry students worked on their own during some lessons whilst she taught the higher tier students. This is a similar approach to Paul who also responds to this type of situation with differentiated provision, whereas Barry responds by teaching higher *and* lower tier work to the *entire* group. She also views timetabling as a constraint on entering Year 11 Triple Award students for different tiers of papers in the three sciences – students can only be entered for the same tier of papers in all three science disciplines.

She considers the practice of entering Triple Award students for Single Award Science as useful for predicting tier allocations. This view is shared with Barry but not Paul. This early entry of students for Single Award is viewed by Cathy as motivational for some but not for others, *'students attach more importance to revision for an external exam, they really learn that* [Year 10 work] *more thoroughly'*, but *'some of them tend to say, well, I've already got my science, I only need a C in science – I got it in Year 10 so it doesn't matter in Year 11'*. Paul and Barry do not comment on the impact of early entry of students for Single Award on their motivation for subsequent scientific study.

Cathy says that the School changed from WJEC to NEAB when the National Curriculum was introduced because the moderation of practical work by NEAB *'seemed to be slightly more in favour of the students than WJEC'*. However, in her view the examination papers of the two groups were and continue to be *'of equal standing'* and *'neither is easier or more difficult than another'*. Like Paul and unlike Barry, she does not express concerns about NEAB's moderation of coursework being incomparable across the science subjects.

When asked whether she views one science subject as being more demanding than another at GCSE, she comments that students think this is so but in her view such discrepancies are only due to different skill requirements. She sees chemistry as demanding some of the skills of both physics and biology '*because there's both aspects - the descriptive parts* [of biology] *and the mathematical parts* [of physics]'. She believes students find biology difficult at GCSE because it requires students to express themselves in English. This source of difficulty was not identified by Barry. My technical investigation showed biology as the science subject with the most positive

correlation value when paired with English. In common with Paul and Barry, Cathy believes students view physics as difficult because of its mathematical demands. My technical investigation showed physics as the science subject with the most positive correlation value when paired with mathematics Cathy believes some students find chemistry difficult because '*they can't get their head around the different names of things that they've got to remember, chemical elements and so on*'. Cathy considers examination papers for the same science subject differ in their skill demands across the different GCSE examining groups and therefore has views on examination comparability. For example, she views WJEC as requiring more continuous prose writing on its chemistry papers than those of NEAB. Thus in common with Paul and Barry, she identifies students' literacy and numeracy skills as impacting on their view of science subject difficulty and an issue when comparing different GCSE groups' examination papers.

She prefers a GCSE system in which all students take a common core examination in chemistry and then '*have the opportunity* [to take] *an extension* [paper]' to provide access to the top two grades because that is *'a far fairer way of assessing them* [students] *without having all these agonizing decisions, well will they get the grade just above so do we put them in for the higher* [tier] *or won't they'*. This reflects Paul's view that tiering is a source of invalidity as it relies on teachers' judgements tracking back to KS3 SATs and if these are not *'right'*, students '*might be mis-entered'*. When discussing tiering issues Cathy observes that the *three* tiers for students' GCSE entry are problematic from a students' viewpoint. In her opinion girls in particular find it difficult to decide whether to enter for a middle or higher tier when they are, according to Cathy, *'borderline'*. She sees this as evidence of girls' lack of confidence in their ability and cites such girls as often being very able and achieving high grades in their A levels and at university. Paul holds a similar view of girls. My technical findings show proportionally two thirds more boys than girls in my WJEC populations which were selected to contain only students who had been entered for higher tier in all three science subjects. As this relative proportion of the two sexes is not shown in the students opting for Triple Award science in School 1, it is tempting to say that this situation reflects a trend to enter proportionally more boys than girls for higher tier. Research into GCSE mathematics tier entries shows this occurs largely because of perceptions of girls' lack of confidence TIMSS (2003). Cathy believes boys are generally not motivated by the mock GCSE

examinations to work hard – but girls are. She takes account of this by treating boys' mock GCSE examination scores as disproportionately lower than those of the girls when making tier entry decisions. She mediates practice to benefit boys based on her subjective view of 'boys' affective responses, thus ignoring the school's policy in relation to the scores. Paul and Barry do not share this practice.

6.2 School / Arena 2

School 2 is an 11-18 comprehensive, co-educational school formed from a secondary modern school some thirty years previous in a new town developed from several small well-established communities. The School has a small number of teachers who taught there when it was a secondary modern school. The town has two other co-educational comprehensive schools, both larger and with better local reputations for their GCSE and A level results than School 2. The data on grouping is largely commented on by Clive, Head of Science Department, with some comments from Betty, the biology teacher.

Year 7 grouping rationale

This is similar to School 1. Students' KS2 SATs' results for the core subjects are used to allocate some 250 Year 7 students to registration groups so that a wide range of average SAT levels are included in each group. These groups are referred to as 'mixed ability'. However, students with the lowest KS2 SAT results and with a Statement of Education Need (SEN) are allocated to a registration group known as SEN. The number of SEN registration groups varies year to year depending on the number of students with a Statement. All Year 7 registration groups are deliberately kept to a maximum of 30 students: the SEN registration groups usually consist of approximately 20 students.

Students are taught science and all other subjects in their Year 7 registration groups. All groups follow the same science course from the commercial scheme, 'Science Now', as for School 1, with the SEN groups' work being pitched *'at a lower level'* (Betty). The science departments' teachers like those in School 1, jointly chose this scheme some five years ago because it is referenced to SAT levels which enable comparisons of students' achievements throughout KS 3. The scheme was also introduced for the SEN groups. This was due to pressure from other colleagues on the Head of Department, Clive, because they wished to use the scheme's work that is

differentiated for SEN students and referenced to SAT levels to justify their recommendations of SEN students' group movements. No reference was made to teachers' adapting or re-writing these tests for the SEN groups of students as in School 1. Lower tiers of tests from 'Science Now' are used for the SEN groups but all 'mixed ability' groups of students sit the same tiers of tests.

Each Year 7 group's science teacher teaches the biology, chemistry and physics aspects of the course regardless of their own science specialism. Movement of students between the 'mixed ability' and SEN registration groups occurs to '*some*' (Betty) extent in both directions during Year 7. Another teacher in this School (a friend, not an interviewee) says it is less than five students out of 250 moving in either direction in any year. This movement only occurs when a student's progress in *each* of the core subjects, English, mathematics and science, indicate it would be appropriate because a student's registration group dictates the whole of their timetable.

Year 8 and Year 9 Grouping Rationale

The outcomes of the 'Science Now' tests and tests in English and mathematics are used to move students between 'mixed ability' and SEN registration groups at the end of Year 7. Otherwise, Year 8 students are taught their science and other subjects in the same registration groups as in Year 7. Uptake of a second foreign language by any student does not change their registration group or impact on their science teaching group arrangements in Year 8, unlike School 1.

The amount of science teaching time increases from four hours in Year 7 to six hours in Year 8. This allows two hours teaching time for each of biology, chemistry and physics as Year 8 groups are taught by three science teachers who teach their specialist subject component of the 'Science Now' course. These arrangements apply to 'mixed ability' and SEN registration groups and the work is differentiated to meet the needs of all students. However, timetabling constraints mean that some registration groups have two not three science specialist teachers, with one of the two teachers teaching outside of their specialism. As in Year 7, science tests taken from the 'Science Now' course are used at the end of topics throughout Year 8. At the end of Year 8 students sit the same examination written by the science teachers using questions drawn from past SAT papers (Betty). All 'mixed ability' registration groups sit the same examination. The SEN registration groups take a modified version with less demanding questions. The results from these

examination papers and tests are collated, equated to national SAT levels and reported to parents at the end of Year 8.

These results are used to allocate students to bands in Year 9 and to identify a group of students with the next lowest average SAT levels to those of the SEN students. This group, plus the SEN students, form a 'lower' band. The rest of the students, approximately two thirds, form an 'upper' band. *'Very few'* (Clive) Year 8 SEN registration group students are moved to the Year 9 top band of students and vice versa. The school has encountered difficulties when drawing up timetables for the two bands of students in Year 9 due to the limited accommodation and number of subject teachers. The number of students attaining the SAT levels 5-7 has meant that the 'upper' band has been too large in recent years for their timetabling needs to be met simultaneously. The school has circumvented this difficulty by identifying about thirty students from this number with the lowest average SAT levels, and timetabled them as a separate group of students called the 'mixed ability transition' group as a part of the 'lower' band. This represents a fine level of differentiation in response to group size constraints.

The 'upper' band of students is given timetabled provision for six teaching groups in each subject. The School gives its subject department heads freedom to allocate these students to teaching groups in any way they wish. The Head of Science, Clive, adopts a setting arrangement. He rank orders the marks of these students' end of Year 8 science examination and uses this information to allocate students to sets in the two Year 9 bands. Table 6.1 represents the Year 9 complex banding and setting arrangements and their naming within School 2.

Table 6.1 Year 9 Banding and Science Set Arrangements – School 2

Band	**Sets**	**KS3 Science SAT Level Preparation**
'Upper' Band	2 'top' sets	6-8
	4 'middle' sets	5-7
Lower Band	1 'mixed ability transition group' (the 'bottom' set)	3-6
	2 parallel groups	2-3

As in School 1, no movement of students *between* bands occurs during Year 9 except in very rare circumstances. In school 2 this specifically occurs when a student in the 'upper' band

fails to reach level 5 in the core subjects. Again as in School 1, movement of students between sets within bands in science is also rare as the science staff aim to have *'got the setting right from the end of Year 8'* (Clive). Any such movement occurs half way through Year 9 as a result of the outcomes of the mock SAT tests held in January. Student numbers in sets and class management issues do not restrict such movements – moving a student up to a higher set does not automatically require the simultaneous movement of another student down to a lower set. If there are set movements, it tends to be from a 'top' to 'middle' set in the 'upper' band because a student is not performing at levels 5-7 rather than students improving their performance and moving up to the 'top' sets. Therefore, a student's KS3 SAT tier entry at the end of Year 9 is largely decided by their Year 8 school test and examination results.

These results and those from the Year 9 mock Science SATs indicate a student's performance in science overall, namely an average of a student's performance in the biology, chemistry and physics components of the science course. No account is taken of differential performance in biology, chemistry and physics. Consequently, not all students have access to achievement in KS3 SAT levels that are necessarily commensurate with their ability in the different science disciplines. As all KS3 Science SAT papers cover all three science subjects within each paper, there is no means of teachers entering students for the science subjects at different SAT levels.

Year 10 GCSE Grouping Rationale

Students choose their Year 10 option courses and timetabling arrangements are in place for the next school year in March of Year 9, well before the KS3 Science SATs are taken. Advice given by teachers to students and their parents regarding science subject choices is largely based on the results of the school's mock Science SATs held in January.

Only two sets of the 'upper' band in Year 9 are entered for the highest tier of papers, SAT levels 6-8. From these 60 or so students one Triple Award group of about 30 students is formed in Year 10. Attaining level 7 or 8 automatically gives a student access to this group. The majority of Triple Award students have level 7 but some have level 6. Students in the 'middle' band of Year 9, all of whom are being prepared for the KS3 SAT levels 3-6 tier of papers, may also achieve level 6 in their mock SATs. It is rare for such a student to be placed in a Triple Award course at the

beginning of Year 10. Students' attitudes are also taken into account when allocating students to Triple and Double Award courses in Year 10. A student's commitment to working consistently hard is used to decide which of the number of students with level 6 should be allocated to the single group of the Triple Award course. Here, hard work mediates achievement levels.

Therefore, entry to Triple Award science courses in Year 10 is largely predetermined by students' allocation to Year 9 'top' sets, which in turn is largely dictated by their Year 8 registration group placement, both of which are determined by their overall performance in all of the core subjects, not just science. Any Year 9 student wishing to follow a career in science is routinely advised by the science department to follow a Triple Award science course in Year 10 and by implication as described above, to achieve level 7 in the mock KS3 science SATs. As in the case of School 1's tertiary college's colleagues, in School 2 Triple Award is routinely viewed by the science teachers as the best preparation for 'A' levels, although students who follow Double Award are still entitled to opt for science 'A' levels at the end of Year 11.

The School uses the national KS3 science SAT results to confirm their decisions regarding students' Year 10 science course placements. In contrast to School 1, these results are *not* broken down into biology, chemistry and physics components. However, a student's differential performance in the science subjects as revealed by school tests and class work throughout Year 9 is a factor in deciding whether Triple Award or Double Award is more suitable for the student. Nevertheless, it only becomes a factor when the differential performance is very marked (Clive). For example, a student who is viewed by science teachers as on line for a grade B in chemistry but grade D in physics and grade C in biology GCSE, is advised to still take the Triple Award course unlike School 1. The rationale is that if taking Double Award, their biology and physics achievements will subsume their ability in chemistry so that they might end up with a grade C or even D: it's viewed as better to get one science with a higher grade. As Head of Science, Clive allocates such students to Triple Award and tells parents and students that studying all three science subjects on the Triple Award course is mandatory: taking the examinations is not, so in the example quoted, the physics examination need not be taken at the end of the course in Year 11.

Science teachers use the mock and the KS3 science SAT results to justify their decisions for parents who wish their child to take Triple Award GCSE Science when a student's SAT

performance is below level 6. Then the School's policy is to present the SAT results to the parents and advise them to enter their child for Double Award GCSE science. Nevertheless, if parents and students still wish to follow the Triple Award course against the science teachers' advice, school policy allows them to do so and is another example of mediation of the assessment process as in School 1. This occurs routinely for about two instances a year and is due to Triple Award being held in higher esteem by parents and teachers than Double Award (Clive). In line with national trends (Institute of Physics, 2006), fewer girls than boys take up Triple Award GCSE science (Peter). The single group of Triple Award course students are prepared for the higher tier of GCSE papers throughout Years 10 and 11 with separate lessons of biology, chemistry and physics being taught by specialists. It is only after students' mock GCSE science examinations in January of Year 11 that entry to a lower tier of papers is contemplated. Students may be entered for a foundation tier for one science subject and higher tier for the others. Students are still taught within the same single group but are given differentiated work that is appropriate for their tier entry.

All remaining students are allocated to nine Double Award teaching sets according to their rank order position in the national KS3 Science SAT results. Each set is timetabled for separate lessons of biology, chemistry and physics with specialist teachers. From the beginning of Year 10 eight of the nine sets are taught work for a particular tier of Double Award science papers: the same tier of biology, chemistry and physics work is taught to any particular set. The sets aiming for higher tier papers are made larger than those aiming for lower tiers to give the benefit of the doubt to students who may be borderline for higher tier entry - but only when they demonstrate a commitment to work, and to aid class management with smaller sets of lower tier students being seen as easier to *'manage'* (Clive). This view and practice was also found in School 1, particularly for Barry. The number of sets being prepared for higher and lower tier papers varies year by year according to the number and 'ability' of the students in Year 10.

Although the eighth set of Year 10 students is taught Double Award science, traditionally very few are entered for the foundation tier of this course. The majority of this set are entered for GCSE Single Award. The vast majority of the ninth set, which is traditionally composed of SEN students, is entered for the Certificate of Educational Achievement (CoEA). Occasionally a student from this set is entered for Single Award and then differentiated work is given to the student rather

than changing the student's set placement, as this would impact on the timetabling of their other subjects.

Timetabling also prevents movement of students between Triple and Double Award courses as these run at different times, as in School 1. Movement of students between the Double Award sets takes place at the end of Year 10 and is largely decided by the outcomes of students' tests and end of Year 10 (summer) examination. Routinely about a fifth of Double Award students are then moved up into sets aiming for higher tier. Approximately the same numbers of students need to be moved down because of class management issues. Movement is only possible between sets that are timetabled at the same time for science to avoid students' other subjects being affected. No movement of Double Award students occurs in Year 11 for the reasons that by then friendship groups are well established and it would be counterproductive for students' motivation and teaching continuity. If a student's progress indicates that entry to a tier different from that of the set is appropriate, the student remains in the same Year 11 set and is given work commensurate with their final tier entry decision. This occurs only for changes from higher to foundation tier. All final tier entry decisions are made in February of Year 11.

Examining Group Choice

WJEC is used for all science examinations. It used to be NEAB but was changed some ten years ago when the current Head of Science Department, Clive, came to School 2. The change arose after Clive's consultation with colleagues and their shared wish to have a local examining group for ease of access to advice and to change from the modular approach of NEAB's syllabuses to the linear approach of WJEC.

6.2.1 The Chemistry Teacher's Perspective and Personal Response: Clive

Clive is in his fifties and a chemistry graduate. He came to School 2 as Head of Science some ten years previous to this interview.

Clive has significant concerns about the use of test results and their *comparability*. He is dissatisfied with the School's current use of KS2 SAT core subject outcomes as the primary indicator of Year 7 students' allocation to registration groups. He desires more information about students' abilities. He would prefer to test all Year 7 students with CATs (Cognitive Ability Tests): he believes these would provide more useful information about students' abilities than the

KS2 SATs. CATs tests are predicated on there being abilities that can be identified as numerical reasoning, verbal reasoning and non-verbal reasoning. Scores on these tests are used to profile students: a high numerical reasoning score predicts ability in science and mathematics, high verbal scores predict ability in language, and a high non-verbal score predicts high general intelligence (www.The British Psychological Society). The results may also be analysed to predict KS3 SAT levels in the core subjects and GCSE grades in a range of subjects (this occurs in the Yellis sytem, Year 11 Information System). Therefore Clive views IQ as being fixed and students as varying in their disposition to learn science. He differs in his view of what is valued knowledge from the teachers responsible for allocating Year 7 students to teaching groups. As Head of Science Department he has the authority to introduce CATs testing for all Year 7 students during their initial science lessons. This practice will begin next school year. He plans to use the outcomes as a baseline for students' progress through the School and has directed his science teachers to conduct the tests and use the outcomes in this way. Thus at the science department arena level Clive is using his authority to mediate the policy that holds at the School arena level of relying on KS2 SAT outcomes as the indicator of Year 7 students' ability.

In his view the School's current policies for the assessment of Years 7-9 provide insufficient information to track students' progress effectively. Clive questions whether the measures of students' achievements used for deciding which teaching groups and tiers of SAT papers students are entered for are comparable. First he is concerned about *what* is measured; he wants this to be 'ability' not science achievement. Second in the absence of his preferred measure of 'ability', he believes it necessary to have a variety of types of assessment information for each student in order to make valid judgements of their relative potential. In his view his staff's class work, homework and test activities are a rich source of information about students' progress and their achievements relative to each other but he has no current means of efficiently collating this information. Consequently, he has asked his teacher in charge of physics, Peter, whom he regards as very IT literate, to devise a computerised database for science staff to record these types of information. He plans to use this information from next year to track students' progress from the beginning of Year 7 to the end of their Year 11. The performance information will be equated to SAT levels.

He feels that Year 10 Double Award science students are not moved often enough between sets (currently, once at the end of Year 10) to reflect their progress and allocate them to an appropriate tier of entry. His rationale is that collecting the necessary data for informing more frequent movements is problematic. He views the creation of the database described above as overcoming this problem because *'you just press a button and the computer will churn it out for you with all the average marks – you are not sitting down trying to work it out and you are not having to go to three people* [each student's biology, chemistry and physics teachers] *to get the data'*. Clive is clearly dissatisfied with his current arrangements for moving Year 10 students between sets but feels overburdened by setting up an alternative system. From next school year, he will use the database to inform decisions about Year 10 set movement that will then occur after one term as well as at the end of Year 10. He sees the advantage of this increased movement as resulting in sets containing students who are more closely matched in their 'ability' and reducing the degree to which he needs to differentiate work within each set. In this way he is like Paul in School 1, who for the same reason tries to get the groups *'right'*.

Clive failed to convince the school's previous Head Teacher to introduce setting in Year 9. On his appointment one year previous to this interview the new Head Teacher allowed subject departments to set or group students in any way they wished within the banding system, largely due to Clive's advocacy. Thus what has been described above about Year 9 sets is a new setting system for science. Clive feels justified in his arguments for setting as the best preparation of students for national assessments: *'this is the first year we have done it* [setting] *and certainly our SAT results seem to bear out that we have got it right because the students all achieved the levels that we thought they should achieve'*. One could argue that the situation is self-fulfilling, that consciously or unconsciously, the sets are taught so they achieve what the teachers hope they will achieve by the new setting system (Hanson, 2000). Nevertheless, he is frustrated by not being able to timetable the 'lower' as well as the 'upper' band of sets at the same time to enable free movement of students between all sets as their progress changes. He recalls there being three or four students in the mixed ability group of the lower band who *'could have gone into the top band's sets but because they are in that form, they had to remain where they were because they are timetabled at different times'*.

He strongly disagrees with the current school practice of using students' performance in all of the core subjects to decide their Year 9 band placement. A students' mathematics, English and science teachers must all agree on his / her banding allocation '*which is not good because a student who is good at English might not necessarily be good at science or mathematics, so it* [banding allocation] *is a compromise'*. His concern is that students have differential abilities in these subjects and are prevented from following courses that offer opportunities to match those abilities. His view of comparability encompasses the notion of 'gradeness', that for example a grade A is equivalent in all subjects, but that different subjects demand different types of skills which are varyingly possessed by students. He recalls '*a student where we* [science department staff] *fought and fought to get this student moved up* [into a higher band] ... *they* [senior school management team] *would not move him and he lost interest and he was a good scientist, so we lost that particular student'*. As Head of Science Department he feels he can only make his views known to the senior management as he has no authority to replace the banding system with his preferred system of allocating any student to Year 9 sets according to their 'ability' for the specific subject. However, he feels he *is* able to respond to the SEN Year 10 students according to their different abilities in science. He routinely prepares and enters the more able of these students for Single Award Science GCSE as well as CoEA so that they *'get a grade out of it both ways'*. He views the small number of students in the timetabled Year 10 SEN group as enabling him to devote sufficient time for meeting their learning needs – differentiation is not a problem.

Clive views the size of sets as a key influence on class management and therefore on the allocation of students to Year 10 Double Award sets. This reflects teachers' views in School 1. He allocates *'more able students in larger groups* [sets] *because they work'* and *'smaller numbers as you go down* [sets]*'*. This is how Clive responds to the need to differentiate work *within* sets. He views social rather than behavioural factors as very important in determining whether to move students between sets in Year 11. By then friendship groups are well established and in his view it is better to prepare a student for a tier of papers different to those of the set, than to move the student to a different set so close to the GCSE examinations. In this way Clive circumvents the constraints of the Year 9 banding system that prevent him from moving students from band to band according to their progress.

After joining School 2 and teaching the NEAB modular GCSE science course Clive felt like his science teaching colleagues, '*a bit fed up ... with 16 modules* [to teach] *and exams every eight weeks'*. He views the modularity as resulting in *'rogue results'* in that students obtained grades higher than expected by science staff. This then led to students thinking '*that they could go on and do 'A' level* [even though] – *they did not have the ability to do it'*. Clive prefers a linear course because it requires students to perform across a larger body of work than modular courses and this is better preparation for 'A' Level. Comparability is thus a key concern for Clive. He has changed from modular NEAB to linear WJEC courses, as has School 1. Consequently, he mediates students' GCSE grades by changing the type of syllabus. He does not consider WJEC as *'easy'* for students to obtain high grades and recalls that WJEC had a reputation for being *'harder'* than most GCSE examining groups – presumably because of its linearity. He believes all of the GCSE groups *'are now levelling out'* due to *'the fact that they are all competing with each other'*. For Clive important factors in choosing an examining group include non-ambiguity of question wording, clarity of diagrams, equitable coverage of the syllabus by examination questions and appropriately differentiated tiers of papers. WJEC is seen to comply with these factors and is also valued by Clive as being *'easier to access'* at its Cardiff base for advice and when *'things go wrong'* than Harrogate-based NEAB.

Clive views biology, chemistry and physics as *'not more difficult than each other'* but inherently different because they *'have different criteria which they are testing and therefore students of different ability will perform differently on them'*. He views biology as *'more descriptive and* [requiring] *more rote learning'* and that this is why there is a common perception that biology is the easiest science subject. Like Paul, Barry and Cathy in School 1, he views physics as *'very mathematical'*, and although he shares Paul's belief that the mathematical demands of GCSE physics have decreased in recent times, he perceives the mathematical demands as still *'daunting to some students'*. In his opinion, students have always seen chemistry as *'never direct'* in that examination questions may be set on a *'topic they have been taught but it need not necessarily be the actual reaction* [they have encountered before]'. For this reason he believes students have to go through a *'double thought process to get to the answer'*, first to recall chemical patterns and then apply this knowledge to unfamiliar reactions to arrive at answers and '*some*

students cannot make that connection'. This notion of transfer across contexts in items was taken up in my cognitive investigation of the WJEC examinations in the skill category, comprehension and application. I did not find any significant differential weightings in these skills between the science subjects, but my investigation was limited to WJEC 1993 – 1995 examinations. Clive does feel that students achieve grades that are *'fairly level'* in the three science subjects and so like the teachers in School 1, sees comparability in terms of grade equivalence across subjects. He attributes views that any one science is more difficult than another to *'the way we were brought up'*, in the sense of being shaped by other people's perceptions of subject difficulty.

Clive believes there have been significant changes in the demands of chemistry syllabuses and examination questions since the introduction of the National Curriculum, a view not expounded by Cathy in School 1. His most significant concern is the inclusion of geology in the Key Stage 3 and 4 Science Orders and the GCSE chemistry syllabuses. He says, '*my biggest bone of contention is that I am not a geologist. I don't agree with the rock section. I know why it went in there but I would have preferred to see some wet chemistry* [instead of the rock section]. *I think they took far too much out and calculations went with them. ... I disagree with the removal of calculations. I would like to see them back – well not back perhaps to the extent that they were but certainly more of them'*. Clive's concerns with the introduction of geology as a new, too 'difficult' topic on chemistry syllabuses echoes Barry's for biotechnology on biology syllabuses.

Clive believes the changes in the national chemistry syllabuses since the introduction of the National Curriculum have had a detrimental effect on chemistry as a subject and reduced the usefulness of GCSE as preparation for 'A' level chemistry. In particular, he views the decrease in calculation work on GCSE chemistry syllabuses as giving students a false impression of chemistry as a subject. Views of a decrease in calculation work on chemistry examinations in recent years causing chemistry to become 'easier' were expressed by Barry about his son's chemistry assessments. Clive says, *'I am not looking at the subject* [chemistry*] at one particular level, I am looking at the subject overall and I think if you want to give students an idea of what the subject is all about, you have to give it warts and all and I don't think we do, not at GCSE – we have taken a lot of the chemistry out'*. He compensates for this perceived mismatch between GCSE and 'A' level with more attention to calculation work with 'A' level students than he ever recalls being

necessary. He believes his biology and physics colleagues also feel that they have *'problems'* in enabling their students to make the transition from their subjects at GCSE to 'A' level but would not be drawn on this issue.

Clive identified other changes in the 16+ national chemistry examinations across time that included: a greater emphasis on analytical skills in chemistry questions across all tiers of papers and particularly on those for the higher tier, a view shared by Barry for biology examinations; more emphasis on analytical skills in chemistry than in biology and physics; a growing similarity in skill demands for the three science subjects since 1995, which was my finding in my cognitive skill analyses in Chapter 4. As a result, he has decreased the emphasis he places on students recalling information and increased his emphasis on developing their analytical skills. He knows *'examiners* [of GCSE chemistry papers] *are restricted to setting papers with various* [prescribed] *types of skills'* but takes *'with a pinch of salt what the WJEC is saying* [about the papers' skill profiles] in terms of *'what actually comes out on the papers'*.

Like Cathy, Clive regards girls as a whole as working more methodically and more consistently than boys *'up to Year 10'* and that boys *'are more laid back, they don't want to be shown to be swots'*. Arguably, these stereotypical beliefs may shape teacher-student interactions and help to promulgate the differential sex group performances identified in my technical investigation in Chapter 4. By the time of the GCSE examinations, Clive believes *'the able boys will have come up to scratch'*, although there will still be a tendency for *'the boys to underachieve more than the girls'*. Girls are viewed as finding physics harder than boys, a view shared by Paul and Cathy. He would not be drawn to comment on the consequences of these views on his practice or indeed on any other issues relating to gender.

6.2.2 The Physics Teacher's Perspective and Personal Response: Peter

Peter is in his early thirties and a physics graduate. He is the school's teacher in charge of physics and was appointed four years ago.

Like Clive, Peter says that the school's banding arrangements in Year 9 are not conducive to his enabling all students to reach their full potential in science. In particular, like Clive, he finds the range of abilities in the 'mixed ability transition group' in the 'lower' band a challenge for differentiating work. He tries to cover different SAT levels of work but feels some of the more

able students would be better prepared for their KS3 SAT papers if they were placed in the 'middle' sets of the 'upper' band.

Peter supports Clive's view of the need to record more information about students' achievements by electronically logging class work, homework and test marks. He says that *'by and large the best people* [students] *in the year do perform better in exams but there is always the chance that somebody has an off day'*. He shares Clive's view that rather than just using examination marks, an on-going record of students' achievements in science subjects facilitates more frequent and appropriate movements of students between teaching sets aiming for specific national SAT levels of work. Peter is charged with putting this policy into practice. He feels he is prevented from responding to changes in students' attitude to work and academic development by the current practice of retaining those set placements throughout Year 9. Like Clive, Peter is concerned to allow for changes in students' effort to mediate achievement by moving them between tiers. Even with the new database to justify set changes Peter feels that the school's timetabling will prevent students being moved '*between bands*' in Year 9. He aims to continue with his practice of alerting his Head of Department to his concerns about particular students but feels that he in turn has little power to change matters and that they have to accept the limitations of banding and its timetabling arrangements.

Peter routinely experiences pressure from parents to enter students for GCSE physics courses that in his view are beyond students' capabilities. Like Clive he believes such parents view Triple Award as having *'more esteem'* than Double Award and *'offering more opportunities for later career choices'*. He responds to such pressure in the same way as Clive, and as teachers in School 1 do, by sharing his record of these students' achievements with the parents to justify his views. Students' commitment to work is used by Peter to mediate their course placement: if he feels that a student *'is on the lower borderline of making it into that class* [course] *but will work hard ... they go in'* but if *'they are idle and won't work a lot then they won't go into that class'*. His experience is that in the occasional case where a student insists on taking Triple Award despite his advice, the student struggles with the work, realises the inappropriateness of their choice and then changes to Double Award.

When describing his GCSE tier entry decisions for last year's Years 10 he recalls *'having enough students* [capable] *of higher tier for three sets'* and the fourth set *'for foundation* [tier] *with some student movement between the sets so we* [science staff] *had the right students in the right classes'*. He therefore attempts to teach all students in any particular set the same tier of work. However, timetable constraints and numbers of students mediate this preferred practice. For the current Year 10 he has *'already had discussions with three students in one class* [set] *and ... made a firm commitment to study foundation ... they will have to remain in the same class* [set] *as we have no other class* [set] *to put them in ... if I am teaching something higher* [tier work] *I will give them something else to get on with'*. In his view Years' 10 and 11 teaching groups have too wide a range of student abilities, despite the department's setting arrangements, so that students need to be prepared for different tiers of papers within the same teaching set. Clive shares these views. Peter believes he can only make one response to this situation, the same as Clive, that is to differentiate the work for students sitting different tiers of papers in the same set. This professionally frustrates him. So a view has emerged that the curriculum needs to be tailored to students' measured levels.

Peter has only taught GCSE courses administered by WJEC. He has the authority to change his physics GCSE course from that of WJEC but does not contemplate doing so. In particular, like Clive he values the closeness of WJEC for accessing advice and attending meetings. He values: the *'style'* of questioning adopted by WJEC on their examination papers, which he sees as reducing the ambiguity in the wording of questions, and the papers even coverage of different topics. Students' grades are *'mostly as expected'*, with the physics papers being *'generally consistent in standard'* across the eight years of his use of WJEC physics examinations.

He has definite views of the nature of physics and categorizes students according to whether '*they can or cannot do well*' in physics. Those who *'do well'* are seen to need *'a certain brain ... a logical, analytical, mathematical brain*', which needs to be paired with a willingness *'to sit and learn a few formulae'*. Peter is alone amongst School 2's teachers in identifying specific behaviours as essential for students' success in their subject. Above all, like Paul, Barry and Clive, he sees the greater emphasis on mathematical skills in physics as presenting even the 'top' students with more learning challenges than found in either biology or chemistry. In his view over the past three to four years *'the mathematical ability in children has declined'*. He responds by spending

more time explaining the mathematics in his courses than he used to, like Paul. He also shares Paul's opinion that in recent years the mathematics on GCSE examination papers *'has become more straightforward'* but Peter still finds *'there is a deterioration in the students' ability to solve what I* [he] *would term very basic calculations'.*

He views biology as requiring a capacity to learn facts about things – things about which *'they are already familiar with ... the human body for example and plant life'* of which *'they have some rudimentary knowledge'.* This view about biology requiring more learning of facts than physics and chemistry is not reflected in the biology examinations having proportionally more marks allocated to recall type items than the other sciences in my investigation of examination paper cognitive skill demands. It is the abstract nature of the facts in physics and chemistry that Peter sees as more challenging than biology for students. His views are based on his experience of looking at examination papers, which also reveals in Peter's view, that physics has the least requirement for continuous prose responses. He believes students view physics as the *'hardest'* and biology as the *'easiest'* subject. Unlike Clive he declines to teach the Year 7 science course where each teacher teaches aspects of all three science subjects. His decision is based on feeling uncomfortable trying to teach outside of his own specialist subject and particularly so for biology where he *'wouldn't be able to explain things in a way that would open their minds to things further on* [KS4 work]*'.*

Peter holds strong views about boys' and girls' motivation to succeed on physics courses. Clive did not express similar views for chemistry. Peter sees girls as generally *'outstripping boys'* in all subjects up to the end of Year 9. He is of the opinion that throughout secondary schooling girls are generally *'prepared to try harder'*, *'learn work'* and *'listen better'*, with their listening skills being seen as *'far, far superior to* [those of] *boy'*, and so is similar to Clive and Cathy. Peter considers the inability to listen as a major problem for boys because it inhibits their capacity to follow him when he's working through the development and explanation of physics theories – *'it's very, very difficult to keep the boys concentrated on following the line of the argument'.*

He is dismayed by the relatively few girls opting for Triple Award in his school–*'it just doesn't seem attractive to them'.* He believes more of the more able girls opt for Double rather than Triple Award because they are attracted to taking languages at GCSE and timetabling prevents

Triple Award being taken alongside them. These girls see Double Award as still enabling them to *'choose to come back to do A Levels in science'*. He thinks that nationally, girls are more willing than boys to take a wider spread of subjects at GCSE. Nevertheless, he believes he has had *'as many outstanding girl students as boys'* during his eight years of teaching. Although he is cautious about generalising from his annual small student cohort numbers, he claims girls generally achieve higher grades than boys in Triple Award physics – *'the girls tend to peak on B-C and the boys tend to peak on D-E, every time* [year]. Teachers carry performance patterns in their heads. Peter's view of girls' and boys' grade distributions is not supported by my technical findings for WJEC or SEG populations. This investigation shows little difference in the nature of these two physics' grade distributions and the grade range around which the sex's '*peak*'; in the instance where there is an obvious difference, the boys peak at grades A – B and the girls at B-C (see Chapter 4).

Peter does not experience girls being reluctant to be entered for higher tier rather than foundation tier papers. However, he believes they are reluctant to continue with physics studies at 'A' level. Although he tells his students they can take 'A' level physics if they have followed Double Award Science at GCSE, he emphasises that Triple Award is better preparation. He does not think this advice deters girls from taking up 'A' level physics but offers no explanation for the significantly greater number of boys than girls in his 'A' level classes. So, unlike Paul and Cathy, Peter does not see girls as lacking in confidence or this as a reason for girls' reluctance to take physics' courses – rather he sees the reason lying within physics as a subject.

His views about the *comparability* of science GCSE tiers and their examination papers clearly shape his teaching and his advice to students. He emphasises the effect tiers have on attained grades. He believes *'it is easier to get a grade C on the foundation tier than on the higher tier papers'* in GCSE examinations because *'a lot of the harder content is taken out of the foundation tier'*. As a result, he advises students whom he considers to be borderline for achieving a grade C, to take the foundation (grades C-G) rather than the higher tier papers (grades C-A*). He clearly views the tier system as introducing incomparability. I chose just one tier of papers, the higher tier, for my technical investigation as I wished to identify any incomparability across subjects' severity of grading and wished to screen out as many effects as possible that would make this less clear. Peter's practice of entering 'borderline' students for the foundation tier is one such

effect that tends to produce skewed results. He views the foundation and higher tier papers as allocating proportionally equal numbers of marks to calculation work but the higher tier's calculations as requiring a higher level of computational skill – in effect, differentiation being achieved with mathematical demands. Peter shares the same views as Paul in the following respects. All questions on the foundation tier papers are seen as *'more straightforward'* than those of higher tier, with there being *'generally one step ... a bit of knowledge to handle and out comes the answer'*. However, when comparing the higher tier of GCSE physics papers with 'O' level physics papers he views the former *'as easier'* because *'they tend to take the student there* [to an answer] *in steps'*.

6.2.3 The Biology Teacher's Perspective and Personal Response: Betty

Betty is in her late twenties and a biology graduate. She is the teacher in charge of biology, having been appointed a year previous to this interview.

I gained significantly fewer insights of this teacher's views of school and departmental policies and their influence on her practice than for Clive and Peter, despite the use of prompts. This may be due to her relative inexperience in the teaching profession (five years) but is probably because she has only been at School 2 for a year, as she appeared comfortable with being interviewed and willingly answered questions.

Betty teaches Year 7 students, like Clive and unlike Peter. She stresses that she covers the same science topics with SEN students, the 'mixed ability' groups of Years 7 and the sets in Year 8. Unlike Clive she mainly differentiates such work by varying the demand on students' writing skills. She emphasises the usefulness of the 'Science Now' scheme as a programme of differentiated work for SEN students because it is referenced to SAT levels and enables her to compare these students with mainstream students and justify her recommendations for moving students between teaching groups. Her emphasis of this point could reflect her recent arrival in the school and a wish to validate her decisions for her colleagues.

Betty feels challenged by the requirement to differentiate work for Year 7 students in their 'mixed ability' classes. She welcomes the current setting arrangements for Year 9 science lessons and refers to students by national SAT levels and observes that *'by segregating them off we can now target them for the right level'*. She emphasizes the levels as having meaning like Paul in

School 1 and unlike the other teachers discussed so far. She sees teaching students in sets as giving her more time with the *'able'* students to teach them the topics associated with the higher tier papers. Clive and Peter had stressed the usefulness of setting as providing more time with the '*less*' able/SEN students. She feels justified in her view because the school's KS3 science SAT results have shown an increase in the proportion of students achieving levels 6-7 since setting was introduced although as with the same point raised by Clive, this could be self-fulfilling - the use of levels to group students gives them meaning and brings them into being. As Hanson writes ' tests transform people by assigning them to categories and then they are treated, act and come to think of themselves according to the expectations associated with those categories' (2000, p. 74). Arguably, Betty's interactions with students reinforce the practice of teaching students in sets as effective and this deepens the belief in the meaning of levels.

Unlike Clive and Peter, Betty does not recall having pressure from parents to enter their child for a higher tier or different science course, probably because she has only been in the School for a year. As a relatively new member of staff, she welcomes working collaboratively with Clive and Peter when making decisions relating to Year 9 students' allocation to Triple and Double Award courses.

Although she may yet to experience it, she does not view timetabling as a constraint on students' subject and course choices in Years 10 and 11 – *'it was in my previous school ... not in this school because ... the Deputy Head in charge of timetabling gets in all their options and then does the options list after that'*. She also values the flexibility offered by the school's simultaneous timetabling of six teaching groups for each of these Years: it enables her to set students and prepare each set for a specific tier of papers that she considers appropriate for students' abilities - and to move students between sets in response to changes in their progress. She views *'the number* [of students] *needing to move down* [a set] *more or less the same as the number needing to move up so the group* [set] *size stays approximately the same with about 28 in the top sets and about 25 or lower in sets 5, 6, 7'*. Like Clive and Peter she accepts this as an inevitable timetable constraint.

Betty, like Peter and School 1's tertiary college colleagues, advises her students that Triple Award is better preparation than Double Award science GCSE for continuing with 'A' level science subjects, *'I talk to my classes about what they want to do in Year 10* [and] *I stress that*

those who want to go on to do 'A' level sciences should really look at doing separate sciences for GCSE'. She does not view girls and boys differently in their willingness and capacity to follow biology at Triple Award GCSE but feels that girls are more likely than boys to continue with biology at 'A' level, which is consistent with national trends. Despite prompts, she does not expand on these views and, like Clive, does not talk about the comparability of attaining grades in different tiers of biology GCSE examinations as did Peter for physics.

6.3 School / Arena 3

School 3 is an 11-18 comprehensive, co-educational school formed some thirty years ago in a well-established, affluent community. The school has expanded significantly over the past fifteen years and now draws about half of its students from surrounding rural areas. Closure of a neighbouring grammar school some twenty years ago resulted in its teachers being transferred to School 3. The School has a local reputation for good GCSE and 'A' level results.

Although there is a Head of Science Faculty, all aspects of teaching and assessing students on each of biology, chemistry and physics are the sole responsibility of three separate heads of departments who largely work independently of each other. These three heads of departments are the teachers that I interviewed.

Year 7 Grouping Rationale

Students' KS2 SATs' results for the core subjects are averaged, as in Schools 1 and 2 to allocate Year 7 students to eight registration groups which contain students with a range of KS2 SAT results and are referred to as 'mixed ability'. However, students with a Statement of Educational Need (SEN students), are allocated to only two of the eight registration groups in a similar way to Schools 1 and 2.

As in School 2, Year 7 students are timetabled for all of their subjects in their registration groups which are paired for simultaneous timetabling. Heads of Departments have authority to group or set these paired registration groups for teaching purposes, again as in School 2. Science teachers are therefore able to directly mediate the school policy at an individual level. The teachers responsible for biology and chemistry teach the 'mixed ability' registration groups. However, the teacher in charge of physics allocates Year 7 students to upper and lower sets for each of the paired

registration groups as like Clive, he prefers to teach students in sets. He uses the students' KS2 SAT results in the core subjects to place the students from each pair of registration groups in their sets.

As in School 2, all Year 7 science teaching groups / sets are taught separate lessons of biology, chemistry and physics by specialist teachers. Unlike Schools 1 and 2 which use 'Science Now', no specific commercial scheme is used in any of the three science subjects: the work is devised by the teachers responsible for these subjects and differentiated to meet students' learning needs. Separate biology, chemistry and physics tests are written by teachers and referenced to SAT levels. These tests are applied throughout Year 7. It is only within physics that the results of these tests are used to move students between upper and lower sets, although such movement during Year 7 is rare. In biology and chemistry all students remain in their same registration groups for teaching purposes throughout Year 7.

Year 8 and Year 9 Grouping Rationale

Unlike School 1 and like School 2, Year 8 science grouping is unaffected by the introduction of a second foreign language on the curriculum. Each Head of Department adopts a setting approach for allocating students to science teaching groups in Year 8. As in Year 7, the pairs of Year 8 registration groups are timetabled at the same time by the school and then allocated by these teachers to upper and lower sets. The two Year 7 registration groups containing the statemented students are re-organised. The statemented students are put into two smaller sets and the remaining non-statemented students form a third set. All three teaching sets are taught at the same time to allow movement between sets according to students' progress. Allocation of Year 7 students to each of their Year 8 biology, chemistry and physics sets is based on the results of their class tests and end of Year examinations in these subjects. Students' set allocations may differ for biology, chemistry and physics.

The Head of Physics and the student's physics teacher come to a consensus view of a student's set allocation based on the marks from tests that are given to the students throughout Year 7. Little movement of students between the physics sets occurs in the Year 7 / 8 transition. In chemistry, the students' end of Year 7 chemistry examination marks are put into rank order and used by the Head of Chemistry to allocate students to provisional 'upper', 'lower' and

'statemented' sets for Year 8. This allocation is then discussed with the students' chemistry teachers. They qualify these allocations with reference to the marks obtained on homework and tests throughout the year and in this way obtain a consensus view. The same process is used by the Head of Biology for deciding students' upper and lower biology set allocation. Again it is the marks from students' tests and homework and the end of Year 7 examination that inform set allocation, as in Schools 1 and 2, despite the test items being referenced to levels. Unlike Schools 1 and 2, neither of the biology, chemistry and physics Heads of Department refer to using SAT level indicators in their subjects or to the assessment outcomes of any of the other core subjects when allocating students to their Year 8 science sets. The students' tests and homework marks are not equated to levels and there seems to be far less preoccupation with turning students' assessment outcomes into levels in Years 7 and 8 than in Schools 1 and 2. This implies a strong normative view is held in School 3, whereas a belief in levels implies strong criterion referencing. In School 3 the interviewed science teachers all appear to allocate students to teaching groups based on a view that on a few marks matter, a view in which a bell curve distribution dominates. In School 3 there is no significant difference in the relative numbers of boys and girls across the science subject sets in Years 7 and 8 in contrast to School 1 where girls dominate the top sets.

The three science department heads use a similar approach for reviewing students' set allocation throughout Year 8, with reviews at the end of each term based on the term's test and homework outcomes. Year 8 students predominantly remain in the same science subject sets when moving on to Year 9. By the beginning of Year 9, tests composed of past KS3 SAT questions are used to enable each head of department to make comparative judgements of students' progress in their specialist subject, regardless of students' set allocations. The heads of department decide together which tier of KS3 SAT science paper is appropriate for each student. Preparation for entry to these tiers is still conducted in the separate lessons of biology, chemistry and physics, and in teaching sets that are not necessarily identical in their student composition, although in general they are.

Year 10 GCSE Grouping Rationale

The three heads of department analyse the Year 9's KS3 SATs results for their own subject. This information is used to reach a joint decision about whether students are advised to follow Triple or

Double Award science courses, although school policy is to give students and their parents a free choice over these courses. Students' achievements in the other core subjects' KS3 SATs play no part in the science teachers' course decisions. In general, slightly more boys than girls opt for Triple Award science, which does not reflect the national trend of far fewer girls than boys opting for Triple Award science (Murphy and Whitelegg, 2006).

All Triple Award students are prepared for higher tier entry. The final tier entry decision in all three science subjects is made during February in Year 11 based largely on the outcomes of the 'mock' examinations. The vast majority of students sit the same tier of papers in all three Triple Award science subjects as in Schools 1 and 2. Proportionally more students opt for Double Award science and generally in large enough numbers that cause the Triple Award students to be usually limited to one timetabled group because of staffing limitations. As in Schools 1 and 2 the timetabling of Triple and Double Award courses at different times for the majority of Double Award students prevents movement between these courses. Due to staffing limitations, the larger the number of students opting for science in any year, the greater the chance that at least one Double Award group will consist of students with a wide range of abilities requiring preparation for different tiers of papers. Commonly, student numbers require three Double Award groups, 'top', 'middle' and bottom', to be timetabled.

Students' KS3 SAT science paper marks rather than their achieved SAT levels are used to allocate them to the three Double Award teaching sets. However, students' who achieved levels 6 or 7 on their KS3 SAT science papers are routinely allocated to the 'top' set i.e. the higher tier GCSE entry, as in School 2. Any student considered *'borderline'* (Phil) for this set is allocated to it and *'adjusted later'* (Phil) if needs be to the 'middle' set for the foundation tier. Movement of a student from the 'middle' to the 'bottom' set occurs if they find Double Award too demanding so that they can be prepared for Single Award science.

Movement of students between the Double Award sets takes place at the end of Year 10 largely based on test results throughout the year and the end of Year 10 examination. Only four to five students are moved between sets. In all three schools, approximately the same numbers of students need to be moved up and down between sets because of class management issues – the sets are all considered large at approximately 30 students. No movement of Double Award

students occurs in Year 11 because it would be counterproductive for students' motivation and continuity in teaching, which are viewed necessary for students' progress. If a student's progress indicates that entry to the foundation tier is more appropriate than the higher tier, the student remains in the same set and is given differentiated work commensurate with foundation tier preparation. All final tier entry decisions are made in February of Year 11, as in Schools 1 and 2.

Examining Group Choice

The heads of department have chosen to use WJEC syllabuses and examinations for all of their 16+ and 'A' level science courses since before the GCSE's introduction. They defended their choice of WJEC against pressure from the previous Head Teacher, who wished to change to the Midlands Examining Group (MEG), as he believed this group would produce proportionally more high grades in science subjects at GCSE.

6.3.1 The Biology Teacher's Perspective and Personal Response: Brian

Brian is in his late forties and holds a doctorate in biology. He is Head of Biology and joint Head of Science Faculty and came to School 3 eighteen years before this interview.

Brian has chosen not to set students in Year 7 for two reasons: he considers keeping Year 7 in their 'mixed ability' registration groups gives the students some stability at a time of many changes and he mistrusts the validity of the national KS2 SAT results. He believes that Year 6 students are often coached for the KS2 SATs and that this reduces the validity and reliability of their outcomes as predictors of future achievements in his subject. So Brian is concerned about KS2 SATs as a valid assessment instrument, whereas Clive's concern about these assessments lies in the validity of their use for allocating Year 7 students to groups. However, Brian values setting as potentially *'benefiting more the high ability youngsters'* and for this reason is currently debating with his biology teachers the pros and cons of introducing it for next year's Year 7. He aims to have students in their *'right'* sets by the end of Year 8, by which time he considers students' biology assessments have provided him with sufficient information to do so. He prefers: to have no movement of students between sets in Year 9 to allow continuity in the relationship between teacher and student in the period approaching the science SATs; to allocate students to appropriate sets by the end of Year 8.

Brian sees student numbers in Year 8 biology sets as limiting the possibility of movement of students between the sets, a phenomenon seen in Schools 1 and 2. 'Top' sets are made deliberately larger (31/32) than 'lower' sets (25/26) because of behavioural concerns about lower ability students: in his view, smaller lower sets are more *'manageable'*. If a number of students are moved 'up', according to Brian, a similar number need to be *'moved down'* to maintain class management. Such movements are decided through consultation with a student's biology teacher and other people as necessary, for example parents who are dissatisfied with their son's /daughter's set allocation. Brian ensures students do not have the same biology teacher teaching them for consecutive years. He values exposing students to different teaching approaches and teachers' personalities so that students do not stereotype biology as a subject with a particular teacher and practice. He does not take gender into account when allocating students to Years 8 and 9 sets: he does not recall there ever being any disproportionate number of boys or girls in these sets, unlike the situation in School 1.

Brian experiences difficulties with allocating students who have moved from another school to his biology sets. He does not view KS2 SAT levels in science as valid indicators of a student's 'ability' in biology. He allocates a student to a 'top' set only when the school from where the student has come stresses that this would be appropriate. He favours a gradual change in assessment practice from Year 7 to Year 9. He prefers the frequent short testing arrangements in Year 7 biology lessons gradually changing to longer more formal tests based on past KS3 SAT 'biology' questions in Year 9. He views this practice as not giving Year 7 *'too big a burden initially'* and enabling them *'to carry more* [remember more work] *for exams'*. In that sense he has responded to national assessment requirements by changing his student assessments so that they can become 'acclimatized' to what they will experience in KS3 SATs.

Like Betty, Brian recommends his students take Triple Award if they are contemplating continuing with 'A' level science studies and sees Triple Award science as a necessary precursor for careers in science and medicine. He advises Year 9 students of this when they are choosing their GCSE subject options in Year 9. He feels that Double Award science *'tends to be the poor relation almost if you like, or in this school anyway'*. He does not prevent Double Award science students from taking up 'A' level biology provided they have good grades but is *'concerned about*

their ability to cope at 'A' level, not so much because they're not intelligent but because they may not have covered some of the work'.

Brian creates sets in Year 10 on the basis of students' abilities and the number of students choosing biology in each of the two timetabled options. For example, in the year of this interview, in one option he has '*a group of about 28 youngsters who range from* [those] *who are going to struggle to do foundation right the way through to youngsters who are off to Cambridge'*. In the other option he has *'about 80 youngsters ... and we have two parallel groups* [sets] *and then a less able group* [set] *- this is the decision that we took this year – we won't do this necessarily every year'*. He decided to have two parallel top sets and one lower set, rather than the usual top, middle and lower sets because he judged there to be more than 50 students capable of entry to higher tier GCSE biology work and prefers to prepare students in any particular group for the same tier of papers, in common with all of the teachers in this research. However, he feels this preferred practice is constrained by timetabling and staff limitations. Even the current two top parallel sets are bigger than he would like – *'I think one has 33 and one has 31* [students] ... *that's far too many but we just haven't got the staff to make it go any further, they* [school's senior management team] *won't allow me another teacher ... I think the lower* [set] *is probably middle 20s so there's not much space there and particularly if they're not very able anyway, we can't move more students into that set'*.

Brian feels that in any particular teaching set he always has too wide a spread of 'ability'. He responds by teaching work for the tier of paper that matches the ability of the majority of the students in a particular Year 10 set. In Year 11 he then prepares the students within the same set for different tiers of GCSE papers. This preparation can be delayed until the second term of Year 11 after the mock examinations – *'in my group ... they've all up to now been taught as if they are going for the higher* [tier] ... *after their mock exams ... I'm pretty certain there will be youngsters who will have to do the foundation paper when they get their* [mock examination] *marks'*. In Brian's view more staff to teach more sets of Year 10 and 11 students would alleviate the current onerous need to differentiate work.

In contrast to Barry and Betty, Brian directs his biology teachers to teach all students in Double Award sets the higher tier work from the beginning of Year 10, if in his view, the majority

of these students are capable of the work. If students find they cannot meet the demands of the work, they are moved set but only: if there is room in the set below (not exceeding 27 for lower sets); the set movement takes place within the first term of Year 10. He is reluctant to move students between sets in the latter part of Year 10 and at any time in Year 11 as this disrupts student / teacher relationships and has a destabilising effect on students' learning, a view shared by the teachers in School 2. He sees this to be more significant than his preference to teach to just one tier of paper in any particular set. Movement of students between Triple and Double Award courses is rare and only occurs within the first few weeks of the Year 10 courses. He considers movement between these science courses as undesirable because of their significantly different content. Despite prompts and unlike Phil and Clare, Brian would not be drawn on whether he experienced pressure from parents to enter their sons / daughters for Triple Award rather than Double Award or for higher rather than lower tiers of GCSE papers.

Like Betty, Brian recalls *'there has always been a tendency for biology to be chosen by more girls than boys'* for their national 16+ examinations, although *'that seems to have lessened a lot recently ... that's probably due to the National Curriculum'*. He views the introduction of the National Curriculum as having had a profound effect on the number and nature of students following GCSE and 'A' level biology courses. Before the introduction of the National Curriculum it was School 3's policy for all students to take at least one 16+ national examination science subject and this tended to be biology because it was traditionally regarded as the *'easy option'*, falsely so in Brian's view. More of the more able students tended to opt for chemistry and / or physics. At this time significantly more girls than boys took biology at 16+ and proportionally more girls than boys continued with 'A' level biology. Since the National Curriculum's introduction and courses in science with elements of biology, chemistry and physics became compulsory for all students up to the end of Year 11, approximately equal numbers of boys and girls have followed biology courses at 16+. 'A' level biology courses are also now more likely to have *'slightly more boys'* than girls but the numbers of boys and girls are now *'pretty much the same'*, whereas *'a few years ago there were situations where we had one boy and 30 girls – the National Curriculum has evened things up quite a lot'*. In Brian's view the National Curriculum has led to biology in general being held in higher public esteem.

In contrast to all the other teachers in this research, Brian stated that he's *'not very interested in assessment'* and *'more interested in teaching biology'*. He does not wish to investigate the results of other examining groups in order to change to a group who award higher grades than WJEC, although he knows of other School 3 teachers who have done so. He views taking on another examining group to *'gain more* [higher grades]'as requiring *'an enormous amount of work'* from himself and his colleagues for little if any beneficial return. He believes that all students should sit the same examination and disapproves of other teachers' *'shopping around'* for grades. He aims to teach biology as well as he can and just *'see what they get* [GCSE grade]'. His priority is to enable students to enter 'A' level biology courses *'knowing the subject well'*.

Like Clive, Brian perceives examining groups as having *'come into line'* in recent years, in that they are similar in their *'standard'* of syllabus and examination demands. He recalls his work as Chief Examiner for WJEC's 'A' level practical examinations and moderator of GCSE practical coursework when stating that he '*knows WJEC*' have made attempts to maintain *'standards'* across time and set examination papers that are acceptable to teachers and students. Nevertheless, he identifies two changes in recent years as being detrimental to students and influencing his practice. First he disapproves of WJEC's use of particular content for differentiating tiers of examination papers because it makes the papers *'biased'* and gives students an incorrectly skewed impression of what biology is like at higher levels of study – '*I think they* [WJEC] *find it quite difficult now to ask higher* [tier] *level questions about straight physiology things like the heart or the kidney ... and so they tend to go for those areas like ecology, particularly for the higher tier'*. Second, he disapproves of the change in emphasis in practical coursework away from practical techniques to general skills such as *'looking at planning, obtaining results, analysing results, evaluating'*. He says the assessment of these procedural skills and understanding creates for teachers *'an enormous burden ... particularly for biology'*. He perceives the biology investigations [the practical coursework] as more difficult for students and much more time consuming than those of chemistry or physics. He views the unpredictability of living things and the length of time that it is necessary in order to observe them doing anything that can be measured, for example in photosynthesis, as creating *'unfairness'* between biology, chemistry and physics practical coursework. He believes this means he has less time for teaching the theory of the biology course than his chemistry and

physics counterparts. This situation has influenced his practice in that he now only gives students investigations that focus on one or two general skills at a time. To that extent he has simplified the experiences of his students so that they can get through the required practical coursework activities. So his views on comparability for practical work across the science subjects have changed his practice and this in turn has changed the nature of biology as a subject domain for his students. He views the investigations set by WJEC as inappropriate because the skills they seek to assess can be tested by a written examination paper. He is also dissatisfied with teachers being required to mark their own students' practical coursework: he's not confident that other schools honestly record the actual achievements of their students as he does. Again, he reveals his concerns with validity and comparability. He feels he has no means of influencing these concerns.

When comparing recent biology GCSE papers with their 16+ counterparts of some 15 years ago, he concludes that there has been a decrease in *'recall'* and a corresponding increase in cognitive demand, in *'understanding'* type questions, a view shared by Barry. My technical findings do not substantiate this view across all my populations, but it is for the WJEC 1995 higher tier. I am interviewing teachers some four years after my technical investigation so Barry and Brian's view could be supported by more recent research, although any such work is not in the public domain. The requirement to write in continuous prose has also decreased but Brian believes the ability to do so is now used more on the higher than the foundation tier at GCSE as a *'differentiation tool'*, thus introducing construct irrelevance. He views physics as having *'the most difficult concepts'*, then chemistry followed by biology and for this reason considers his own subject to be the *'easiest'* of the three and physics the most *'difficult'*. This is his view: he believes people in general, including the students, also hold this view. My technical findings only support this view for physics for the 1993 WJEC and SEG 1994 populations. In his view difficult concepts in biology are not encountered until studying at university and then *'they are far more sophisticated than most of what you learn in physics or chemistry, whereas in physics you get to very difficult stages quite early on'* [in secondary schooling]. He justifies his view of science subject difficulty with reference to the national KS3 SATs' results where students generally get lower marks on the physics questions than on either of those of chemistry or biology. Brian also considers, as do Cathy and Peter, that the subjects vary in their skill demands. Like Cathy he

believes that biology has a greater emphasis on recalling facts than either chemistry or physics, a view not supported by my technical findings. Like Paul, Barry, Cathy and Peter he also associates physics with mathematical demands, *'if they're* [students] *not good at maths then they've got a major problem as a far as physics is concerned'.*

He identifies the difference *'a very good physics'* teacher can make to a student's motivation and ability to *'do well in it'*, but still regards the science subjects as inherently different in their *'rigour'*. He refers to students' saying that they like / do not like a subject usually because they like / do not like the teacher of the subject. The degree of influence a teacher can have on a student *'frightens'* him and he refers to his own practice of trying to start lessons *'smiling and laughing and not tired'* as this influences *'how students feel about the subject and how they cope with them* [subjects]'.

6.3.2 The Physics Teacher's Perspective and Personal Response: Phil

Phil is in his early fifties and a physics graduate. He is responsible for all aspects of teaching and assessing physics in the School and is co-Head of Science Faculty with Brian. He came to the school when the neighbouring comprehensive, in which he was Head of Science Department, closed some eight years ago.

Like Clive, Phil prefers to teach students in sets and chooses to set Year 7 students, unlike his biology and chemistry counterparts. He considers students' mathematical skills as being fundamental to learning physics, a view shared by Paul, Barry and Peter. This view influences his practice: unlike Brian he bases his set allocations on KS2 SAT results in all of the core subjects but with particular reference to the mathematics as well as the science results. Students' assessment outcomes in mathematics are also taken into account when he reviews students' set allocations throughout Year 7, particularly when a student is borderline for 'top' set allocation.

He views Year 7 as an important year for refining his initial set allocations and refers to engaging students in a *'diagnostic process'*. He uses the same physics assessments for all Year 7 physics sets throughout the year as a means of *'comparing'* their progress. By the end of Year 7 he has used the rank ordered outcomes of tests used throughout the year to move students between sets for Year 8 but *'just on the margins ... occasionally ... very little change, but some change'*. Like Brian he feels his setting practice is constrained by the numbers of students and available staff;

movement of a student to a higher set usually requires him to move another student to a lower set. He deliberately keeps the lower sets smaller than their upper set counterparts *'to allow for increased student contact* [with their teacher] *...and tuition'*. This suggests that Phil is concerned about the relationship between class size and students' learning rather than behaviour and management issues as was Paul. Like Brian, Phil rarely moves students between sets in either Year 8 *or* Year 9 and does not take gender into account when allocating students to sets. Unlike School 1 but like School 2, in general there is no significant difference in the boy : girl ratio in any of the Years 7 to 9 physics or biology sets.

Phil shares Brian's concerns about the lack of valid information from other schools about students' 'ability' in science subjects to inform his allocation of new students to his physics sets. He admits that *'there will be an occasion when we might have got it wrong* [inappropriate set allocation] *when somebody comes in* [joins the school during Year 8] *and we are working on the Year 8 information* [SAT levels] *from the previous school'*. Brian relies on the previous school's set recommendations to allocate incoming students. Phil also uses this type of information but in contrast to Brian, feels that he has to seek out the information directly from the previous school as the senior teachers within School 3 are not forthcoming with this information. This perception is similar to Barry's about his school's senior management staff on this issue, perhaps reflecting in both schools the existence of two 'communities of practice' (Wenger, 1998) where the expectations of one are not fulfilled by the other. Generally Phil is cautious and unlike Brian tends to allocate incoming students to a lower set until they demonstrate to him or his physics colleagues they can cope with the higher set work.

Phil values the faculty's policy of analysing the KS3 SAT science results to give each student's marks in the physics, chemistry and biology questions. He finds the physics marks useful for qualifying his judgements about students' ability to take various Year 10 physics courses and therefore for qualifying setting allocations based on other information including *'teacher assessment, test results - test scores and so on just in case again it was an off day* [the day KS3SAT tests were taken by the student]. Year 9 students' options mean that generally there are only numbers sufficient for one Year 10 Triple Award group. When numbers warrant two groups, Phil allocates the students to two sets based largely on their physics KS3 SAT marks. He teaches

both sets higher tier work and delays tier entry decisions until February of Year 11, a practice Brian requires of all science staff. Then, if a student is viewed as suitable for foundation tier, he adopts the same practice as Brian by keeping the student in the same set and giving them differentiated work.

Phil allocates students opting for Double Award to three sets but unlike Brian, does not occasionally create parallel 'top' sets and *always* has 'top', 'middle' and 'bottom' sets. He uses the same sources of information about students' abilities for allocating students to these sets as he does for his Triple Award set allocation. In common with Brian's approach, the physics 'top' set is deliberately made larger than the lower sets for two reasons. In common with the teachers in Schools 1 and 2 Phil sees higher set students as *'easier'* to manage than those of lower abilities. He is also keen to give *'borderline'* students the opportunity to be prepared for higher tier work and so access high grades; then he moves such students down to a lower set preparing for foundation tier when their class work and test outcomes indicate it is necessary. These movements, like those for biology, occur by the end of Year 10. Like Brian and School 1 and 2's teachers, Phil values the stability given to students' by remaining with the same teacher in the run up to the external examinations in Year 11.

In common with Paul, Barry, Cathy and Peter, Phil views the difficulties students have with their physics courses, be they Triple or Double Award, as largely stemming from their mathematical skills – *'they* [students] *can select data from information ... do correct substitution* [into equations] *but often getting the answer at the end is for some reason a weakness'*. Like Paul his response is to teach mathematical skills in his physics lessons. In his view girls find the mathematical demands of physics *'a little bit more trying than the boys'*, which he perceives as due to their lack of confidence rather than ability. My technical findings from exploring the relationships between boys' and girls' physics and mathematics GCSE performances do not indicate that girls' relatively poorer performance on physics GCSE is linked to a similar relatively poorer performance on mathematics GCSE, tempting me to suggest that other factors are at play. Indeed Murphy and Whitelegg (2006) suggest that the girls and physics issue is a multivariate problem. Phil finds there is little overall difference in the abilities of boys and girls opting for physics in Year 10. Unlike the national trend (ibid.), in general only *slightly* more boys than girls

opt for Triple Award and 'top' and 'bottom' sets of Triple and Double Award physics courses have approximately equal numbers of boys and girls.

He finds that girls have a *'different attitude'* to boys – *'most of the girls, most of the time want to do it properly. Some of the boys, some of the time, don't want to do it properly'*. He views girls as *'all very neat and careful and meticulous in their approach'* and *'want to get it right'* whereas *'boys will tend to look for shortcuts - do it very well but as quickly as possible'*. This difference in attitude does not result in girls out-performing boys or vice versa in their GCSE physics results. He sees boys *'surging at the end'* [of the course in Year 11] when they *'suddenly put the football away and decide they have got to open some books now ... and the ability is there and they come through'*. His states that although girls generally have less confidence than boys, they are more likely to volunteer that they are having difficulties with physics and are easier to support because of this, whereas boys will *'mask it'*. This reflects Paul's experience of more girls than boys being willing to attend his extra mathematics lessons for those students who think they are having difficulty with physics.

Phil recalls the previous Head Teacher of the school putting pressure on the science staff to change from WJEC as a GCSE examining Group to MEG *'on the grounds that they* [MEG] *were seen to offer more grade As'* in all three science subjects but particularly in chemistry. Phil's response was to analyse the grade results awarded nationally by these two GCSE groups because he questioned the Head Teacher's claim. Indeed Phil emphasises his interest in exploring GCSE examining groups' practices for comparability, unlike Clive and in marked contrast to Brian. He found no significant differences between the groups in the grades they awarded for chemistry and biology. However, WJEC's physics results showed that only 10% of all grades were allocated to grade A compared with 20% for all other GCSE Groups. He recalls reporting these figures to WJEC at the time and attending a meeting of science teachers at WJEC offices, at which he was surprised to be informed that *'people* [physics teachers] *were moving to other boards* [GCSE groups] *but equally important, people* [students in Wales] *doing physics was collapsing from ten thousand down to two thousand'*.

WJEC personnel had used this meeting to inform science teachers that they were changing their practice of only allocating 10% of all grades to grade A in physics. The following year

WJEC's physics results were similar to other GCSE groups in that they too showed 20% of all grades being allocated to grade A. Since that time the number of students entering GCSE Triple Award has steadily risen for WJEC. In Phil's opinion this is due, at least in part, to an improvement in WJEC's science syllabuses, becoming available when required and appropriately detailed. At his previous school he had changed from WJEC to Southern Examining Group (SEG) because WJEC required him to choose *'which paper* [tier] *a student is entered for and that limited their grade range'*. At the time SEG had a system of a common examination with access to grades B-G and an optional extra paper to achieve grade A. Poor performance on the optional paper did not penalise the grade achieved on the common paper. Phil preferred this system but by the time he had moved to School 3, all examining groups had come in line to WJEC's system as required nationally. Like Cathy, Phil views tiers as mediating achievement.

He believes there have been changes in 16+ physics syllabuses and examination papers that regardless of examining group, have resulted in his subject becoming *'gentler – less rigorous'* than when 'O' level was current, that is pre 1988. My technical findings support this view of physics becoming less severely graded from the 1995 sample in my investigation. Like Paul, Phil feels that now examination *'questions are more straightforward, broken down, more accessible'* and *'more structured ... so that kids can see what people are getting at'*. Like Peter he also views the questions as *'less demanding mathematically'*, a view based in part on his practice of giving *'many 'O' level questions from yesteryear'* to current *''A' level students who say "oh is this 'A' level Sir,* [it's] *very demanding" '*. This closely resembles Paul's experience in School 1. Students are no longer required to learn physics formulae, they are provided on the papers, and Phil believes this, together with the removal of some *'difficult'* topics such as geometric optics, has led to the current physics GCSE being *'easier'* than its 'O' level counterpart.

Although Phil prefers a common paper with an optional one for high grades, he likes the three-tier entry system for science GCSE [this applied at the time of my gathering examination results for my quantitative focus on comparability] as it enables him to know *'exactly what* [topics] *was going to come* [on each tier of papers]' and it was *'a lot easier for teachers*, himself included, *'to get kids to pass examinations'*. This is because each tier of papers only covered specific topics. He found it easier to prepare students for their examinations on a more limited amount of work

than is required now in the two tier system (which still applies in 2008). Like Peter, he thinks mathematics is used to differentiate the current two tiers of physics papers – *'questions set on the lower tier tend to be straight forward substitution ... as opposed to manipulating formulae* [on the higher tier] ... *there's more on the higher tier'*. This mirrors Brian's view that the skill of writing in continuous prose is used to differentiate the higher tier of biology GCSE examinations. Phil stresses that this makes him feel he needs *'to practice, practice, practice ... the manipulation of numbers'* with Year 10 students as he prepares them for their GCSE physics examination questions.

He does not experience pressure from parents to enter their child for Triple rather than Double Award or tiers of papers that are higher than he recommends and says *'it seems our judgement is trusted* [by the parents]'. Phil feels there is a tendency for both boys and girls to want to *'play safe'* in physics: students who are judged by him as incapable of gaining a B-A* grade, tend to want to enter the lower tier (grades C-G) providing them with a C grade, rather than risk not getting a C on the tier aimed at A* - C.

He believes that he has to work harder than his biology and chemistry colleagues to motivate students to learn physics because of students' preconceived ideas about it – that it is the *'most rigorous* [science subject in school]', and its concepts are *'abstract'*, views shared by Peter and Barry. Phil's response to this view is to *'translate* [the abstract concepts] *into real events'* and to provide his students with *'a nice secure environment so they are happy in what they are doing and creating an interest'* as *'then it will get easy'*. None of the other teachers interviewed for this research referred to their provision of secure environments for responding to students' insecurities in studying a particular subject. Phil views the decreasing number of 'A' level physics students, graduates and teachers compared with their biology and chemistry counterparts as exacerbating students' willingness to study physics. He is particularly dismayed at the quality of physics trainee teachers *'coming through'*, which *'means people* [the trainee teachers] *are patching up on a subject* [physics] *which they have no in depth feeling for ... who can't see the relationships* [of physics concepts] *in the real world'*. Off tape, Phil recalled the poor quality of physics teachers applying for a teaching post at School 3: they often had a degree in which physics was only a small part or their physics degree was of poor quality.

6.3.3 The Chemistry Teacher's Perspective and Personal Response: Clare

Clare is in her late thirties and a chemistry graduate. She is responsible for all aspects of teaching and assessing chemistry in the school, a post she took up some ten years previous to this interview.

In common with Brain and unlike Phil for their subjects, Clare does not find national KS2 SAT science results useful for predicting students' capacity to learn chemistry. She prefers to use her own assessments and those of her colleagues during Year 7 to make judgements about students' abilities. This is why, like Brian for biology, she organises the teaching of chemistry in 'mixed ability' registration groups rather than sets in Year 7. Setting of students for chemistry lessons occurs in Year 8, allocation to particular sets being largely decided by *'the examination results at the end of Year 7 ... a list is drawn up, top set, lower set and special needs and then at a departmental meeting or some other occasion that would be passed around ... and* [we] *decide if we felt people* [students] *were misplaced'*. Anomalies between these examination results and students' performance on class tests and homework throughout the year in her experience occur *'occasionally, so somebody perhaps who on the basis of the exam results should have been lower set,* [is] *put into the top set'*. It is the teacher's *'knowledge of the student'* that is paramount in set allocation.

Comparability in assessment is a key concern for Clare. She requires chemistry teachers to record students' achievements on assessments that are common to all chemistry students within each year from Year 7 to 9, from tests based largely on past KS3 SAT chemistry type questions to homework activities. She values this system for providing measures of students' achievements on a common scale. Comparisons of assessment outcomes are discussed in terms of marks, rather than SAT levels in accord with the science faculty's arena level practices and reflecting the three teachers' apparently shared belief in norms. She admits to using predominantly recall type questions in the chemistry tests given to students during Years 7 to 9. Her rationale is that these tests, *'rightly or wrongly'*, are *'to make sure that they know their work'*. Therefore, she routinely uses assessment for reinforcing students' factual / rote learning rather than just for making judgements about students' course or tier entry.

Students' set allocations are reviewed at least three times in Year 8 based on assessment information. Set movements occur largely after Christmas of Year 8 and in common with Brian

and Phil, are then only adjusted at the beginning of Year 9, to avoid students having a succession of different teachers. Set movements in Year 9 are rare as Clare believes she has *'got it more or less right by then'* and like Brian and Phil views continuity in the relationship between teacher and student as paramount with the approach of KS3 science SATs. The impact a teacher's personality can have on the quality of a student's learning is a key concern for Clare like Brian. For example, she is currently questioning the wisdom of having moved some Year 8 students to her lower set: having now taught them, her view is that based on their current achievements they should have remained in the upper set and says she is '*beginning to think personality of the teacher* [of the upper set] *is affecting them* [the students in the upper set]'. This view reflects a concern for the quality of teaching affecting access to learning; differences in these are ignored in the technical approach to examination comparability as discussed in Chapters 3 and 4.

All Year 8 and 9 sets follow the same work in the same order. Setting is viewed by Clare as a means of teaching the same work more effectively to students with different abilities. The chemistry teachers work *'closely together without it being too prescriptive'* and have *'departmental meetings* ... [so they] *know roughly where everyone is so promotion and demotion* [across the sets] *can take place quite easily'*. All Year 9 are prepared for the higher tier of KS3 SAT papers as with Brian. However, it is the students in the chemistry department's lower Year 9 sets that are usually entered for the lower tier of KS3 science SAT papers. As for biology and physics, across time there have been no significant differences in the proportion of boys and girls in the Year 9 chemistry sets, or indeed, in any of the chemistry sets for Triple Award and Double Award in Year 10.

Clare does not take account of students' mathematical abilities when advising them about GCSE tier entry – *'the only time I would become aware of what they are doing in maths is perhaps when I come to do an investigation* [Year 10 and 11 GCSE practical coursework]...*some of them will decide to do a time of reaction whereas others will decide to do a true rate and then I need to know what maths* [skills they have]'. Then she provides additional support by explaining the mathematics to the students who are finding it *'a little bit tricky'*. She says that students opting for Triple Award have usually achieved well in mathematics' assessments, for example in their KS3 mathematics SATs, and do not require her to explain the more mathematical aspects of chemistry

found in Triple Award syllabuses. She did not raise mathematics as an issue when discussing Double Award students.

The number of students opting for Double Award in Year 10 determines how many chemistry sets can be timetabled by the school and in turn whether Clare can set them or teach them as one 'mixed ability' group. When sufficient numbers opt for three timetabled Double Award teaching groups, unlike Phil she avoids setting them as 'top', 'middle' and 'lower' sets as she sees *'the only people who have a feel good factor are the top set'* and this in her view is counterproductive for students' motivation. Like Brian she then creates two parallel 'top' sets and a 'lower' set because she thinks *'if you can avoid someone saying "oh I'm in the third set" it's so much better'*. In contrast, when faced with a particularly able group of students opting for Triple Award in sufficient numbers to warrant the school timetabling two teachers to teach them, she allocates them to 'top' and 'lower' sets. Her rationale is that such labels would *'get the A*s out of them* [the 'top' set]'. Therefore, she appears to value setting for motivating able students to learn, as do the teachers in School 2.

She rarely moves students between chemistry sets on the Double Award course because all students are taught higher tier work and in the same order up until February of Year 11 when GCSE tier entry decisions are made. Movement between these sets occurs usually for social reasons such as friendship problems. If a student requires preparation for entry to a tier that is different to the majority of their set counterparts, differentiated work is provided.

Like Phil, Clare has not experienced parental pressure to enter their child for Triple Award rather than Double Award. However, unlike Phil, Clare has experienced parental pressure to enter students to higher rather than foundation tier in both Triple and Double Award. In such circumstances she provides parents with her assessment records for the student and advises them that the evidence shows her decision is appropriate. It's rare that parents do not then follow her advice.

Clare has only ever prepared chemistry students for WJEC examinations. When *'big changes such as GCSE coming in'* have occurred, she has investigated other examining groups' syllabuses but each time has decided to remain with WJEC. Her reasons are that she: has *'experience of WJEC'*; '*knows the examinations lady in chemistry quite well*' and that this is useful

when seeking advice; and she feels she *'can always 'phone them'*. She values the experience she has of their examining practices, of knowing *'the system'*. She has no concerns or comments regarding their chemistry examination papers and views the syllabus as being appropriately reflected in the examination papers. However, she has significant concerns about the national KS3 SAT science questions and the chemistry component of the national curriculum. She says that she feels chemistry has been changed into another subject because of *'all the geology on the chemistry part of the national curriculum'*, a view shared by Clive. She feels powerless to change matters, as indeed she is.

6.4 Reflection

My main interest in this Chapter has been to follow up my findings from Chapter 4 to better understand if comparability is an issue for teachers, what their views are on this and how their practices and arena practices mediate 'gradeness'. A rich picture has emerged with comparability at its heart. I have found significant mediation at both school and departmental arena levels but also, and importantly, at the individual teacher level within schools. This mediation typically reinforces assessment structures and practices and the view of 'ability' and teaching they assume and shapes how schools, teachers, students and parents perceive assessment levels and science award structures.

In the next chapter I examine these shaping processes within school departments and their sources and their implications for students' access to and learning in science. I look at what disrupts these and how this impacts on what 'gradeness' means.

Chapter 7

Looking across arenas and teachers

7.1 Approach

In this chapter I use the interviews in Chapter 6 and relate them where appropriate to my technical findings to examine how schools, departments and individuals respond to national assessment structures and practices to understand how these responses influence what 'gradeness' means in terms of common currency. In so doing I reflect on what the interviews opened up about assessment processes in secondary school science departments to show how structures and beliefs shape arena practices and teachers' practice, the consequences for students' access to science, and the consequent validity of assessments of their science 'achievements'.

The method of constant comparison was used to look for similarities and differences between schools and their practices, teachers and their practices and the beliefs that inform them, and between science subjects. Adopting Wenger's notion of *'constellations of interconnected practices'* (1998, p. 127) discussed in Chapter 5, an arena can be a school or science department. The former embraces the latter and exemplifies practices that are not in the control of the department, in the same way that departmental practices are outside of the control of teachers unless they are part of the senior management with responsibility for setting up structures and processes.

Whole school policies, structures and practices that I refer to as *school arena effects* are shown by this study to include those concerned with grouping of students, reporting of achievement, timetabling and student numbers. These school policies, structures and practices are established to address influences beyond the school, that is: the requirements of national assessment and its associated reporting; national examination award structures; and school resources in terms of staffing availability and time. Schools' responses can reinforce or disrupt these influences. Similarly, school practices mediate and impact on teachers.

The study shows that science departmental structures and practices that I refer to as *departmental arena effects,* include science teaching schedules / courses of work, choice of GCSE examining group, and student teaching group and setting arrangements. School practices influence

departmental practices and both influence individual teacher's practice within settings. Teachers in the same department may be constrained by the same arena structures and practices, for example the department's policy of using KS3 SATs to allocate students to specific KS4 courses. The interviews show that the interactions between the different arena and setting practices lead to consequences not only for students' GCSE science tier entry but for their learning opportunities throughout secondary school and beyond.

Individually the teachers respond in different ways to arena level structures and practices as they orchestrate their settings. For example, one teacher will teach all students the same work at KS4 and delay tier entry decisions until the second half of Year 11; another will allocate students to specific tiers of papers for teaching purposes from the beginning of Year 10. This research shows that teachers' decisions when entering their students for GCSE science examination papers vary according to different kinds of influences operating across time, and operating at least from secondary school entry. Although limited in the amount of data collected from the interviews, it is clear how teachers in orchestrating their settings within arenas (school and department) are both using assessment *and* being constrained by assessment in ways that vary depending on the school and their individual perspectives, values and beliefs. In my discussion below I refer to these variations in practice as *setting effects* as they emanate from the teachers within their settings. They influence how teachers respond to departmental, school, and outside influences, for example parental pressure to enter students for higher tier examination papers.

In what follows I use the sociocultural notion of planes of analysis and levels of mediation (Rogoff, 1995) described in Chapter 5, which are shown to work in both directions from the social level down and back, to structure the discussion.

7.2 The mediation of school practices: assessment and curriculum pathways

7.2.1 Treating assessment measures as surrogates of 'ability'

All three schools are required by Government to report to parents annually on their children's progress in relation to SAT levels of achievement for Years 7-9. This national policy requirement, therefore, shapes how schools interact with parents about their children, which in turn shapes the discourse within

schools about how students and their achievements are conceptualised and described. This was evident in the practice shared by all three schools who used an average of a student's KS2 SATs results in English, mathematics and science to compare students and allocate them to groups, be they in bands (School 1) or mixed ability (Schools 2 and 3). This practice at school arena level validates the use of SATs, demonstrates a belief in SAT levels as valid measures of both students' achievement and potential and the comparison of these. This validity is challenged not least by one believing as Brian (School 3) that some students are coached for their KS2 SATs and that the test outcomes reflect the effectiveness of that coaching, which will vary from student to student.

7.2.2 Grouping by 'ability'

Table 7.1 summarises the schools' grouping arrangements for Years 7 to 11, students' movements between these groups and the timing of KS3 SAT and GCSE tier decisions. The schools differ significantly in how they organise students for teaching purposes in Years 7-9. School 1 uses mixed ability grouping within a banding system which changes in nature from Year 7 to Year 8 due to the introduction of German on the timetable. Nevertheless, there is grouping by SAT achievement from entry. School 2 uses mixed ability grouping throughout Years 7 and 8 with these groups being put into bands in Year 9 at which point the school's measures, referenced to SAT levels, are used to group by achievement as a surrogate for 'ability'. School 3 has the least complex organisation by pairing mixed ability groups for timetabling purposes and giving departments freedom to organise the 'pairs' as they wish throughout Years 7-9. This practice at department level results in some 'pairs' being retained as mixed ability groups and others being changed into sets.

However, for all three schools it can be argued that a student's GCSE science course and examination tier entry in Year 11 is set up by their teaching group placements in Years 7 to 9. The practice in all three schools is to rank order students using the average of their KS2 SAT core subject results to allocate them to Year 7 'mixed ability' groups on entry. School 1 timetables these groups at different times as 'upper', 'middle' and 'lower' bands. Movement between these bands and groups is only on the agreement of all core subject teachers. As each band is prepared for specific tiers of KS3 science work from the beginning of Year 8, one could argue that in School 1 a student is 'locked onto'

Table 7.1 Curriculum and Assessment Pathways

KEY: Band refers to a whole school structure for differentiating students in any Year whereby students are put in broad ability groups for curriculum / timetable provision usually based on their core subject achievements or curriculum need, for example taking a second foreign language.
Set refers to a departmental structure for differentiating students into teaching groups usually based on their science achievements.
SEN = Statement of Educational Need TA = Triple Award DA = Double Award B = Biology C = Chemistry P = Physics

	School 1	School 2	School 3
Year 7	Banding at school level based on average core subject achievement. Movement between bands once in January, based on all core subject achievements, and not uncommon.	Mixed ability groups and SEN groups at school level based on average core subject achievement. Movement is rare and only between SEN and mixed ability groups based on all core subject achievements.	Mixed ability B and C groups based on average core subject achievement. No movement between these groups. Sets in P based on all core subject achievements and an emphasis on science and mathematics KS2 SAT results. Rare movement between these sets is based on physics' and mathematics' class work.
Year 8	New bands and sets within them are created at school level. **KS3 tier chosen at the beginning of Year 8.** Movement between sets within a band in January is rare and based on science achievement.	Mixed ability and SEN groups as in Year 7. **KS3 tier chosen end of Year 8** Movement is rare between SEN and mixed ability groups and is based on all core subject achievements.	Paired groups are formed into sets for each of B, C and P. **KS3 tier chosen beginning of Year 8.** Movement is rare between the paired sets.
Year 9	Same sets as in Year 8. Movement between sets is rare.	Banding on a school level. Sets within bands for science. Rare movement between bands based on all core subject achievements, or between sets based on science achievement.	Same sets as Year 8. Movement is rare.
Year 10	**TA Biology and Physics tiers are chosen at the beginning of Year 10. TA Chemistry and DA tiers are chosen end of Year 10.** No movement between DA and TA groups. Some movement between sets within TA and DA occurs end of Year 10.	Only one TA group. No movement between TA and DA groups. Some movement between sets within DA groups based on science achievement end of Year 10. **TA and DA tiers are chosen end of Year 10.**	**TA tier chosen beginning of Year 10.** No movement between TA and DA sets. No movement between TA sets. Some movement between DA sets end of Year 10. **DA tier chosen end of Year 10**
Year 11	No movement	No movement	No movement

a particular science assessment and curriculum pathway from entry as a result of arena practices at both school and departmental levels. Table 7.1 shows that the same argument applies to the physics students in School 3 and to a lesser extent its biology and chemistry students.

7.2.3 Disrupting views of 'ability' through arena practices

Unlike Schools 1 and 3, teachers at School 2 delay associating particular tiers of science SAT levels to each student until the beginning of Year 9. This practice and the teaching of the same work to all students in Years 7 and 8 arguably mediates the potentially negative influence of differential success at the age of 11 impacting on students' access to science in secondary schools. Other research (Boaler, 2002) suggests that this practice might provide School 2's students with more flexibility to develop and avoid the self-fulfilling scenario created by the tier-associated banding and setting arrangements in Schools 1 and 3. In School 3 it is the marks not the levels that are used to rank order students suggesting that performance is not considered in relation to subject criteria reflected in broad levels of achievement but rather is norm-referenced in relation to peer progress. This is evidence of how apparently small changes in arena practices can significantly alter the intentions within national level structures and practices. The school also recognises that departmental views of subjects can alter views about the need for, and value of, grouping by ability. Therefore the decision to leave grouping choices at that level gives wider opportunities for departmental practices to mediate the assessment process. The physics department's view of the nature of physics knowledge and its reliance on a sequenced subject like mathematics might explain in part why setting was introduced in Year 7 but not for Biology or Chemistry.

Nevertheless, student movement between grouping arrangements in Years 9 - 11 in all three schools is a rare event. Clearly the practices in School 1, and for some students in School 3, provide students, who may have had very different opportunities to learn science, with very little scope to disrupt labels of incompetence handed to them at a young age, and before they have opportunities to experience some of the cultural tools associated with science by engaging in laboratory-based work. All of the interviewed teachers favour movement between their student grouping arrangements in

Years 7 to 10, be they sets or mixed ability, and desire a greater frequency of such movement. Consequently, school practices constrain departmental and teachers' mediation.

Students in schools where setting by SAT level achievement is not implemented have more opportunity to achieve rather than more potential. The schools' grouping practices change the opportunities to learn made available to students and 'gradeness', which is predicated on an assumption of equal access (Chapter 3), is undermined.

7.2.4 Timetabling and access

The argument above about the mediation of potential achievement by arena practices is further supported by School 1's re-organisation of students into two new bands for all subjects in Year 8, based on whether or not a student takes German. Students in each band are rank ordered for placement into science teaching sets, which again are prepared for specific tiers of science KS3 work. The 'top' sets in both bands are not comparable in terms of their achieved KS2 science SAT level profiles. It is more likely that a German band student is prepared for entry to a higher tier KS3 SAT level paper than a non-German band student when they have achieved the same KS2 science SAT level. This is because of timetable constraints and student numbers: the German band has insufficient timetabled 'space' to create sets that differentiate students by their science achievements and teachers mediate by giving students 'the benefit of the doubt' and preparing them for higher tier papers. Therefore choosing to study German provides greater access to science for some students in School 1. The introduction of a second foreign language in Schools 2 and 3 does not impact on students' registration or other subject teaching groupings in the same way as in School 1.

Timetable constraints are experienced by all of the teachers in the three schools and influence the movements of students between teaching groups and sets that are commensurate with their progress in science. The banding systems in Schools 1 and 2 result in students in different bands being timetabled for lessons in the same subject at different times. A student is prevented from moving across bands for one particular subject because of the timetable implications for their other subjects. The arena practice in both schools requires a student to perform equally well or badly in all three core subjects to warrant band movement. As students in different bands are prepared for different tiers of

SAT level papers, this provides further support for my argument that students are 'locked onto' assessment and associated curriculum pathways.

Another influence outside of the school is recruitment. In all three schools the number of students in any Year is another key school arena influence on whether students can be moved between groups and sets and prepared for different tiers of work. This is apparent in the Triple and Double Award groups and sets in Years 10-11 as well as in the Years 7-9 under discussion.

Timetabling is shown to constrain students' access to learning and preparation for tiers of KS3 SAT levels / GCSE papers. Again 'gradeness', which is predicated on an assumption of equal access (Chapter 3), is undermined.

7.3 The mediation of departmental practices

The assessment and curriculum pathways, and indeed what 'ability' students need to have to be on them, may differ from school to school because of the practice at science departmental level. For example the departmental practice in School 1 provides a separate Year 9 set for SAT levels 6-8 preparation alongside those for SAT Levels 5-7 and 3-6. At School 2 the timetable is very constrained and the department cannot mediate this school effect by providing a sufficient number of sets to enable students to be prepared in this way. Therefore, the more staffing resources, the more the opportunities are for access to science. Consequently curricula experiences are mediated at the school and departmental level and vary between schools. The interaction of achievement and opportunity to learn is not taken into account in technical analyses of comparability or assumptions about 'gradeness'.

The influence of national assessment scheme requirements and schools' reporting policies mediate departmental practices and dominate and shape teachers' thinking about their students. At science department level teachers in Schools 1 and 2 chose the 'Science Now' scheme largely because its work programmes and tests are tailored to particular SAT levels. This scheme has a co-ordinated science approach and for this reason is not chosen by School 3's teachers with their separate science teaching approach. The scheme is valued in Schools 1 and 2 for reporting students' progress in SAT levels as required nationally and for making SAT level tier entry decisions because it provides performance data using the same scale as in the national assessment arrangements. It also

differentiates for SEN groups and allows comparisons between these groups and others according to the teachers. Therefore the department's need to compare students dictates students ' curricular experiences.

School 3's teachers want their students' science achievements related to SAT levels but this only becomes important to them in Year 9 with the approach of the national KS3 SATs. They use students' achievements in marks on past KS3 SAT papers to guide their tier entry decisions. In School 3 the model of normative peer referenced assessment is distinctly different to that intended by the national assessment model.

Movement of students between teaching groups be they set or mixed ability is seen as desirable by some of the study's teachers for ensuring that such groups are prepared for the same particular tier of KS3 SAT papers (or tier of GCSE papers as discussed later). This appears to reflect their view of needing to offer a very targeted curriculum experience as they believe strongly that 'ability' fixes what can be achieved, as exemplified by Phil in School 3. For others such as Brian in School 3, it is less important as they have a more fluid view of students' development.

7.4 The mediation of teachers' practice and beliefs

Across the interviews various beliefs and theories emerged which influenced teachers' responses to assessment structures and practices at school and departmental level. These included:

- human achievement as flexible and open to teaching or fixed by IQ or innate cognitive skills;
- assessment levels and marks as valid measures of ability and potential;
- the cognitive skill demands and 'difficulty' of science subjects;
- the representation of subject knowledge in national examinations and views of valued knowledge;
- gender : difficulty interaction.

These are now discussed but positioned under headings that capture different aspects of their nature.

7.4.1 Views of human achievement

Students' SAT levels of achievement are commonly not valued by the teachers for placing students in Year 7 science groups, be they sets, bands or mixed ability groups. They are also not valued for

informing decisions about science set placement when a student moves from one school to another. Each of the three schools' Head of Science Department view the average of students' KS2 SATs as an invalid predictor of future scientific achievement. Underlying this view are teachers' views of human potential as being open to teaching or fixed by IQ or innate cognitive abilites such as those measured by Cognitive Ability Tests (CATs) (Clive in School 2).

Brian (School 3) challenges the validity of the KS2 SAT outcomes, believing that Year 6 students are coached to such an extent for the KS2 SATs that the outcomes are invalid and unreliable indicators of ability. This suggests that Brian is aware that the SATS do not measure ability and potential but opportunity to learn and so can be influenced by teaching. Like Barry (School 1), Brian appears to want to give students the chance to achieve and shows he is prepared to change from mixed ability to setting in Year 8 when he thinks this arrangement is timely for producing more effective learning. Brian sees learning as dependent on the structuring resources made available, hence the richer and more challenging the opportunities made available to students, the more likely they are to achieve their potential. In contrast Paul (School 1) and Clive (School 2) believe that ability determines what can be learned and therefore see themselves as responsible for handing over and determining the knowledge that is appropriate for particular students, otherwise students are faced with unfair challenges and teaching is undermined because the curriculum offered is not targeted. Cathy (School 1), Peter and Betty (School 2), Clare and Phil (School 3) also share this view. Brian's view also influences the practices in the departmental arena: he advises other teachers to allow for unknown, unanticipated potential. Therefore he believes in progression that is not fixed by a notion of innate ability. So his view of mind mediates the view assumed in the national assessment procedures and in the arena grouping practices.

For other teachers the practice of using SAT levels of achievement for placing students in Year 7 science groups is criticised because an average of the KS2 SAT scores in the core subjects is seen as an invalid surrogate measure of ability: they want an approach that accepts students' different strengths being related to subjects rather than IQ or general ability. Such views lead to differences in teachers' orchestration of their settings within the same arena with consequences for students' learning and

assessment. This is exemplified in School 3 with Phil using students' mathematics and physics achievements to guide their different physics curricular experiences in sets in Year 7 – 9, whilst Brian provides the same type of curricular experiences for all students in mixed ability groups for biology.

7.4.2 The paradox of teacher assessment discourse

Paradoxically in all of the interviews, but to a lesser degree for School 3, teachers of Years 7-9 students talk about their schemes of work, their test arrangements, students' achievements and refer to teaching groups and sets in terms of *SAT levels* in line with school and departmental practice and practice beyond the school. Teachers talk about SAT levels rather than about particular skills or types of knowledge being developed, and preparation for particular tiers of SAT level papers rather than stages in understanding scientific concepts. Arguably this is because of the national requirement for them to measure and report their students' achievements in terms of SAT levels for comparability – and school accountability purposes, and they are so preoccupied with this requirement, it dominates their discourse.

Teachers in the three schools validate their KS3 SAT tier entry decisions for students in Year 9. They all do this by using the results from a 'mock' KS3 science SAT assessment consisting of past papers a term before the national SATs. Across the schools teachers commonly use these outcomes to justify their course recommendations and tier entry decisions when challenged by parents. Indeed, the SAT levels in general are valued by all the teachers for providing a comparative measure of students' achievements in science and informing decisions about students' movements between sets and bands. SAT levels are viewed by all of the teachers as the 'lowest common denominator' for comparing students' achievements and the level allocated to a student at KS3 is a primary influence on teachers' GCSE science course and tier entry decisions.

The teachers' use of SAT levels indicates their view of how achievement and ability can be validly represented. So within teachers' discourse there is corresponding acceptance of the *meaning* of SAT levels between the school arena practices and the teachers' in settings. It is very powerful how much agreement there is in the teachers' interviews about the meaning of these levels. SAT level is a social construction / representation that has come to be accepted in the last 15 years. Its acceptance

influences all three schools' practices and teachers' practice. For example Cathy and Paul, Clive and Peter, and Clare and Phil were all concerned to improve the accuracy of their assessments of students SAT levels in order to create homogeneous 'ability' groups to get the curriculum and its assessment just right (differentiated) for the students and themselves. This assumes that the assessments are rigorous and valid measures and that the levels have common currency. Barry and Brian have different views of learning that lead them to be less concerned with moving students between groups for the same reasons as the other teachers.

7.4.3 Arena constraints and teachers' views of achievement

The way that teachers deal with arena level practices provides further insights into their beliefs about mind and achievement and what they understand is essential to enable learning. Teachers viewed timetabling and student numbers as constraints which mediated how students were allocated to particular science courses and moved between teaching groups. These constraints have differing impacts depending on why teachers want to move students. Cathy, Paul, Betty, Clive, Peter, Clare and Phil have their practice of creating homogeneous SAT level and GSCE tier groupings significantly constrained by these school arena effects. Barry and Brian who want to keep opportunities for learning open to all students had to respond to these school arena effects by treating groups as homogeneous (they teach to the same SAT level or GCSE tier) and risk making it too difficult for some. Therefore these two arena effects constrain all of the teachers' preferred practice.

Brian and his colleague, Clare, mediate their school's grouping practice in the same way: they leave Year 7 in their allocated 'mixed ability' groups whilst they gather assessment outcomes from students' homeworks and class tests. These outcomes are then used to set students in Year 8. Colleague Phil does not question the validity of assessment *per se* as does Brian. However, he shares the same view about valuing his own assessments more than the KS2 SATs, as predictors of future achievement - but responds differently to the school arena effect. He cross-references the students' KS2 SAT science outcomes to their KS2 SAT mathematics outcomes to allocate Year 7 students to sets. This practice reflects his view of the importance of mathematics for learning physics and his

acceptance of the validity of both assessments as measures of ability. Then, throughout Year 7, Phil modifies these set allocations in line with his own assessments of students' science ability.

In School 1 the Head of Science mediates the school's Year 7 allocation policy as follows. He rank orders the students' KS2 *science* SAT results and uses them to allocate the bands of students to Year 7 science teaching groups which are mixed ability in terms of these results rather than the average of all core subjects. He therefore views students as having different types of cognitive abilities and selects the outcomes of the assessment of the subject that is most relevant in his view. He clearly values the outcomes of the KS2 science SATs as predictors of potential achievement in science. Paul and Barry are both influenced by this departmental practice but respond differently according to their own views of ability and learning. Paul has strong views of students as learners who need to be given a differentiated curriculum so that the challenges are within their capabilities. This is why he seeks movement of students between sets whenever there is disparity. Barry seems to have a broader view of students' potential and likes to keep opportunities open. He does not differentiate the curriculum and seeks to give students challenges. He sees peer-peer interactions as influential on learning and moves students between sets when these interactions are counterproductive. He also uses movement between sets as a motivator for students whatever their ability which reflects his view that students learn from being challenged and by being given affirmative feedback. His stress on peer interaction also suggests that he recognises that learning is supported by dialogue and that other learners serve as learning resources. These characteristics of his pedagogy suggest that he sees learners as responsible for their learning and active constructors in the learning process; teachers are the providers of the resources and guidance to help them progress.

Clive, the Head of Science in School 2, does not view KS2science SAT outcomes as predictors of potential achievement in science. He believes in generic ability and seeks to implement a surrogate measure of it within School 2. Clive was in the process of mediating the school's Year 7 practice by introducing Cognitive Ability Tests (CATs) to provide him with his *'own base level'* of students' ability that in his view is a more reliable indicator than KS2 SATs. The CATs' outcomes are to be used to modify students' group and set allocations in Year 7 -9. Like Paul he encourages movements

between teaching groups / sets so as to get the curriculum differentiation right, that is, the knowledge targeted at the assessed level of the student. This suggests that Clive shares Paul's view of an ability 'ceiling' on students' potential. Although Peter agrees with Clive's departmental practice of monitoring students' progress, he does so for different reasons. He is more like Barry in his view of students' potential. He also sees examination performance as situated and from his comment about students having *'off days'*, mediated by social factors. His comment that physics needs a *'certain type of brain'* suggests he has a broader view of students' potential than provided by the notion of IQ. He may also be reflecting on the nature of physics and the way of looking at the world associated with it and what is valued in this.

In common with Peter, Clive values the use of his own assessments of students' ability, for example his class tests and homework, for students' set allocations but mistrusts other teachers' assessments. Brian and Phil prefer teacher's views of a student's science ability generated from classwork and homework rather than SAT levels for allocating students to sets when they transfer from other schools. A fellow professional's recommendation of a student's ability is valued and acted on.

School 3's arena practice of giving subject teachers freedom to organise Year 7-9's timetabled pairs of 'mixed ability' registration groups in their preferred manner is used most by the physics teacher, Phil. Here too I would argue that students are *locked onto* an assessment path. Phil allocates students from each pair of registration groups to sets from the beginning of Year 7. He prepares these sets for specific tiers of science SAT papers, getting them in the *'correct'* set for tier preparation by the end of Year 7. He sustains these groupings through Years 8 and 9 for KS3 SAT entry and describes movements between sets as *'rare'*. Phil monitors the match between his tier entry decisions and students' KS3 science SAT outcomes but does not reflect on the potential for them to be a 'self-fulfilling' influence on the student. Teachers like Brian and Barry, also mediate the arena effect of inadequate numbers of timetabled sets by teaching work for both tiers to all students in a set and delaying tier entry decisions until late in Year 11. Barry adopts this practice and finds that the disadvantage is that there is as a tendency for lower ability students to lose interest.

Thus timetabling is usually a constraining school arena influence on students' access to preparation for tiers of SAT papers and GCSE examinations that match their needs. Teachers' differing mediation of this arena effect results in differing access to learning opportunities for students. Comparability which assumes students achieved grades reflect a background of equal access is again shown to be undermined.

7.4.4 Beliefs about students' ability and behaviour - teacher : student relationship

In the interviews it was clear that there was an association between 'ability' and behavioural problems for many of the teachers. In all three schools the number of students in any year is a key school arena influence on whether students can be moved between sets and prepared for different tiers of work. Teachers mediate this effect by making 'top' teaching groups / sets larger than their 'bottom' counterparts. This practice stems from teachers' view of needing to give as many students as possible the opportunity to be prepared for higher tier entry. With the exception of Paul and Phil, it is also because 'lower' group / set students are generally viewed by the teachers as being more challenging in their behaviour and easier to manage in smaller numbers rather than that smaller numbers allow for more one-to-one contact time, which Barry valued for the lower sets. An equal number of students are moved between groups / sets to avoid them becoming overly large and unmanageable (low 30s is the maximum acceptable number) in all three schools, except in Year 9 in School 3. This exception is because this particular Year 9 has lower numbers of students than usual but with the same timetable provision. Barry although stating that lower ability groups *'have more behavioural problems'* seems to have a more nuanced view of the teacher : student relationship arguing that behaviour and potential are strongly linked. He also considers that some students, perceived by other teachers as having behaviour problems, as *'not being any trouble at all'* in his classes. Further, he argues that students should have different exposure to teachers and teaching approaches so they do not come to associate a subject with a particular teacher and practice. Clare too noted the significance of the teacher : student relationship when she commented that students who had been moved to a lower set that she taught were in her view misallocated.

School 2's teachers delay students' tier allocations until the end of Year 10, and appear to value students' attitudes to working hard more than the teachers in Schools 1 and 3. They use it as a rationale to mediate tier entry decisions for Triple Award allocation significantly more than the teachers in either School 1 or 3. A similarity in practice amongst all of the teachers is their retention of Triple Award and Double Award students in the same sets throughout Year 11 and for the same reason – teachers value continuity in teacher: student contact in the run up to the GCSE examinations. All of the teachers prefer to prepare students for different tiers within the same set rather than move students between sets being prepared for particular tiers in Year 11. What this practice suggests is that all teachers recognise the significance of the teacher : student relationship as another factor that can mediate their opportunity to learn and hence their achieved grades.

7.4.5 Representations of science and subject difficulty

Only School 3's teachers use any biology, chemistry and physics specific achievement information in their decisions for national KS3 science SAT tier entry. This may be because only School 3 has decided at a department level to teach the sciences as separate subjects from Year 7 and has this type of data readily available. Nevertheless, the students' three science specialist teachers in School 3 still come to a consensus view of students 'ability' when deciding students' KS3 SAT tier entry. Students' performance in the separate science subjects only appears to become significant for the teachers in all three schools during the transition from Year 9 to Year 10 when GCSE subject choices are being made. Up until this point, it is a student's average performance on all three science subjects that determines their national KS3 SAT level entry. Consequently, national assessment arrangements and school policies prevent public recognition of a student's comparative excellent performance in any particular science subject – it's subsumed within an average SAT level result for all three science subjects. This makes the KS3 SAT science results reported as levels serve as blunt instruments for comparing students' science achievements – and is why this is mediated by teachers in schools 1 and 3 analysing the results into their component separate science marks. It also means that students have no feedback to inform future decisions about GCSE science course choices which impacts on the possibilities available to them post 16.

The agreed departmental practice in Schools 1 and 3 of analysing the science KS3 SAT results for each student's separate biology, chemistry and physics component marks is so that they can use these marks to validate their decisions on students' Triple and Double Award course and tier allocations for the beginning of Year 10. This is to mediate the school arena practice of both schools planning timetables for the next academic year before the KS3 SATs are taken - teachers seek reassurance from the analysed component marks about the validity of their decisions. Teachers in School 2 also use the science KS3 SAT results to validate their course and tier decisions but in terms of the overall science result with no reference to any biology, chemistry and physics marks.

School 2's practice of mixed ability grouping extending through Years 7 and 8 should allow for wider access to the curriculum and potential for learning than Schools 1 and 3. However, once in Year 9, School 2's school and departmental practices in terms of banding and setting, mean that they offer fewer opportunities than Schools 1 and 3, for example accessing GCSE qualifications in all three sciences. Clive as Head of Science Department is pivotal in this respect. He has the most structured view of knowledge and operates a hierarchy that sees modular GCSE science as less like 'A' level and for this reason chooses linear GCSE science syllabuses, a practice that is imposed upon and thus influences the other teachers. He sees Triple Award as being of greater '*value*' than Double Award to the extent that he advises students to take Triple Award rather than Double Award even when they are likely not to achieve well in one of the three science subjects. He then advocates students not taking the examination in their weakest Triple Award subject. As Clive is Head of Department his practice in this respect and his beliefs about the nature of knowledge extends as an arena effect to influence other teachers' practice.

School 1 is unique amongst the three schools in its departmental practice of entering Year 10 Triple Award students for Single Award GCSE Science and using the achievements as an indicator of students' potential achievement in tier entry decisions for Triple Award. This ex-grammar school used to enter its more able students for GCE 'O' levels a year earlier than the normal age – but rather for acknowledging accelerated learning than as a predictor of future achievement and allocating access to other 16+ courses. Indeed all of the teachers' individually appear to view KS3 SATs, Single Award,

Double Award and Triple Award as a hierarchy of awards related to students' ability rather than future needs which is how they were intended. The National Curriculum in science and the statutory requirement for all students to continue their study of science until they are 16 years of age is predicated on an entitlement model of curriculum, but the national assessment structures and teachers and schools take up of these filter students according to their abilities for allocation to particular courses and tiers of work. This in turn alters how they are positioned to continue their education in science post 16. Teachers are therefore shown to reach their decisions regarding students' KS4 science courses in differing ways in which the nature of the information used to inform their decisions varies, for example specific separate science achievements garnered from KS3 SAT component mark analysis or from classwork and homeworks or from the whole science KS3 SAT level allocations or from a combination of any of these. This variety undermines comparability of access to science courses and students' achievements.

Of all the interviewed teachers only Phil comments on examining group tier organisation and how it affects his practice but in so doing he reinforces the view that assessment shapes teachers' practice and as a consequence, the access that students have to science. Phil views the three-tier entry system for science GCSE, which applied at the time of my gathering examination results for my quantitative investigation of comparability, as enabling him to know *'exactly what* [topics] *was going to come* [on each tier of papers]' because each tier of papers only covered specific topics. Then, he found it easier to prepare students for their examinations on a more limited amount of work than with the current two-tier system. Although the tiers are each taken to be a valid representation of science differentiated by 'difficulty', it is clear that the teachers recognise that they differ in broader terms and that the meaning of a grade achieved on different papers must therefore be questionable. This is clear in the way that post 16 a grade B from an intermediate tier in maths is considered to be incomparable to a grade B from the higher tier and students with the latter are more likely to be accepted onto 'A' level study (Murphy and Whitelegg, 2006). This raises further questions about how assessment artefacts change the representations of the subject and therefore the meaning of grades and their comparability.

7.4.6 Choice of examining groups

Examining groups are a key influence mediating school and departmental practices. The choice of examining group is therefore one way in which at departmental level teachers can mediate the representation of valued knowledge for their students. Syllabi vary, although these variations are not considered significant in altering the meaning, and hence the validity of, grades in relation to different syllabuses in technical analyses of comparability. Teachers' practices showed that they disagreed with this.

All three schools have a policy of giving the Head of Science Department (HOScD) freedom of choice with respect to GCSE examining groups and syllabuses. Only Phil refers to pressure from his Head Teacher to choose one particular examining group. This was because the Head Teacher believed a certain group was likely to award proportionally more 'top' grades – pressure which Phil was free to resist as a Head of Department by arguing he was working within the school arena practice of freedom of choice. All of the HOScDs consult their science colleagues when choosing an examining group. The reasons for teachers choosing a particular group include the availability and quality of syllabus and support materials in line with curriculum changes such as the introduction of practical coursework, ease of access to group personnel, 'fair' moderation of practical coursework, examinations that reflect the content of syllabuses, and familiarity with a group's systems and personnel.

The HOScD's choice of group is mainly influenced by the type of available syllabus. School 1's HOScD largely chose his group (NEAB) because he wished to prepare his students for a modular course and the local group, WJEC, at this time only provided linear courses. Conversely, Clive as HOScD in School 2 moved from NEAB because he wished to prepare his students for the linear courses available from WJEC. Students have no choice in whether they pursue linear courses with end of course examinations on the entire course content or modular courses with regular assessment of discreet course content throughout the period of their study. Arguably students will differ in how they respond to these two assessment approaches and one can anticipate that this practice will impact on

students' achievements. Cathy (School 1) views modular tests as advantaging the weaker students who had trouble with memorising.

Brain, co-HOScD with Phil, is singularly disinterested in exploring examining groups for differential awarding practices. This appears to be due to his experiences as a WJEC 'A' level practical biology examiner and GCSE moderator making him feel confident about this group's practices. Certainly School 3 has used WJEC for more than 15 years, despite Phil's concerns about the low rate of grade As in WJEC's GCSE physics. Phil believes bringing his concern to WJEC's attention resulted in an increase in awarded grade As, so his concern has been allayed. If bringing his concerns to WJEC's attention really did change this group's awarding practices as it appears, teachers certainly can mediate assessment by their actions and thus individual agency can influence practices outside of the school arena. A further example of mediation from within schools on structures and practices beyond them is evident in School 1's response to the concern about the stringency of their moderator of chemistry coursework for Double Award. At a department level they mediated the moderator's influence by changing from submitting coursework for biology, chemistry and physics to just biology and physics and subsequently had their students' Double Award coursework marks being treated more leniently during moderation. This is an example of how institutional level mediation can resist the impact on students' achievements of external influences.

7.4.7 Views of subject

Peter articulates a view of the nature of learning physics as requiring a *'certain brain ... logical, analytical and mathematical'* and in common with Phil identifies physics as being more challenging than biology for his students because of its abstract nature. Peter does not offer information about how this impacts on his practice but Phil responds by translating the abstract concepts into real events. Across time Peter and Phil also view the physics curriculum as having become less rigorous in terms of its concepts; further qualification of this view was not provided. Interestingly my technical findings indicated that from the 1995 examinations the study's populations achieved higher mean grades in physics than in 1994 and 1993 and there was a simultaneous increase in the weighting of recall of knowledge type questions. Apart from this perception of physics becoming less rigorous, all three

physics teachers identify and focus on the mathematical demands of their subject – this is a shared, significant concern. Paul, Peter and Phil all view the mathematical demands of physics as the reason why students report that they find it to be the most 'difficult' subject to learn of all the sciences. The chemistry and biology teachers share this view. For example Brian believes this explains why his students obtain proportionally lower marks for physics than either biology or chemistry in their science KS3 SATs; Barry uses this view to justify his belief that it is harder to get a *'good grade'* in physics than in chemistry or biology. Other than my technical study finding a high positive correlation between physics and mathematics achieved GCSE grades – and the most positive values compared with mathematics and chemistry and biology pairings, I have no other findings to illuminate this issue.

All of the physics teachers view the mathematical abilities of their students as having deteriorated across time. Apart from Paul citing problems with the quality of mathematics teaching in his school, no other insights are offered to justify this view. My technical findings found no significant deterioration in the mean grades of my populations in GCSE mathematics but I view the dependability of these findings as limited as described in Chapter 4. They also view the mathematical demands of physics national assessments at 16+ as having become less rigorous across time. Examination questions are viewed as having become less computationally demanding and more structured so that students are now led through calculations. For these reasons physics 'O' level GCE is equated in its intellectual rigour with the current 'A' level. Examining groups are also viewed as using computational requirements to differentiate foundation from higher tier GCSE physics papers.

The physics teachers seek to mediate these external structural influences embedded in assessment artefacts. Their common practice is to provide their students with mathematics teaching within their physics curriculum time to supplement that provided in the mathematics curriculum in their schools. In addition Paul responds by running special short courses of mathematics outside of the School 1's timetable and by producing 'mathematics for physics' booklets for his students. In this way Paul is changing the opportunities to learn and extending access to physics by recognising mathematics as a key tool of the subject. All three physics teachers take particular account of students' computational skills when deciding students' GCSE tier allocations and specifically when a student is

considered borderline for higher tier entry. Phil also takes Year 7 students' mathematics KS2 SAT results into account when allocating them to his physics sets. These types of mediation are unique to the physics teachers and suggest that views of mathematics performance may mediate access to physics through both teachers' and students' actions.

The chemistry teachers, Clive and Clare, consider that the nature of the subject has changed substantially since the introduction of the National Curriculum. Interestingly, my technical findings indicated that from the 1995 examinations when syllabuses based on the National Curriculum were examined at GCSE for the first time, the study's populations achieved relatively lower mean grades in chemistry than in 1994 and 1993. Clive and Clare regret the introduction of geology to the chemistry in the National Science Curriculum at Key Stages 3 and 4. They both feel powerless to mediate this situation. Both teachers refer to students' mathematical skills. Claire feels she does not need to take these skills into account when making course and tier decisions and rarely feels she needs to teach such skills within her chemistry lessons. Clive believes that since the introduction of the National Curriculum there has been a significant decrease in calculation work based on chemical concepts on GCSE chemistry papers, a view shared by Barry comparing his son's chemistry examinations with those from previous years. In Clive's view, this misrepresents the subject and leads students to have a false view of chemistry and to think that they are capable of 'A' level chemistry, which according to Clive has retained calculations based on chemical concepts. Clive's mediation includes advising students with high GSCE grades of the difficulties of the GCSE : 'A' level transition and teaching more mathematical skills at 'A' level than ever before.

All three chemistry teachers are unique amongst the teachers in referring to their students needing to learn and recall large amounts of 'knowledge'. I note that my technical study did not find any consistent differences in the recall of weighting knowledge for the three science subjects based on my PGCE students' views. Claire tests students on their work as a means of motivating them to remember chemical facts: she uses assessment to mediate learning and reinforces a particular view of the subject. Clive views chemistry as being more demanding of students' analytical skills than either physics or biology on GCSE papers. Again I note that my technical findings from the investigation of

the WJEC GCSE Triple Award 1993, 1994 and 1995 papers do not support this view, although I have questioned the dependability of these technical findings in Chapter 5. Clive's view of the subject influences his practice: he has increased the emphasis he places on analytical questions in his chemistry assessments. Students' achievements on these questions influence Clive's course and tier entry decisions and demonstrates further how teacher's beliefs mediate the potential for students' to gain access to science and opportunities to learn.

Biology teachers Barry and Brian view their subject as having become more demanding since the introduction of the National Curriculum, a view supported by the technical findings showing relatively lower mean grade scores for the populations in my study from the 1995 examinations. In their view GCSE biology syllabuses are now overloaded with content so that they are now challenged to complete the work in time for the KS4 examinations. Their response is to teach the content at a faster rate than they desire and this influences the opportunities they have to interact with students which in turn impacts on their learning. Brian also mediates the course content challenge by only setting practical investigations that focus on one or two skills at a time. So through his practice to meet assessment demands, Brian alters the nature of practical work and the representation of biology as a consequence. This may ensure that students' achieve their grades but suggests that in different settings with other biology teachers in other schools a good grade in coursework will vary in what it means raising a further issue about 'gradeness'. Barry and Brian also view the GCSE examination papers as now including more data handling and analytical questions than before with a simultaneous shift away from recall type questions. Both of these shifts in valued knowledge align with approaches that boys are typically associated with and have more familiarity with outside of school. Barry and Brian agree that in recent years more difficult concepts have been included on GCSE biology papers including biotechnology, which is viewed as particularly problematic due to syllabuses lacking clarification of its requirements and the lack of textbooks. Introducing new areas into a domain necessarily alters the representation of the subject and therefore what tiers and their associated grades mean. The lack of training and resources to support teaching these new components will have an impact on student performance and may account for fluctuations over time in apparent severity of grading. Thus quite

different explanations could account for shifts in grade profiles and challenge the technical approach to examination comparability which does not take them into account.

All of the biology teachers reflect these changes in their teaching and assessment of students. They argue that these changes warrant movement of student and public opinion away from regarding biology as an 'easy' science subject. In Brian's view this is happening already and cites an increase in his male students wishing to take 'A' level biology as evidence of biology now being held in higher esteem. It might also reflect the shifts in the domain which boys might find particularly engaging. However, the chemistry and physics teachers still regard biology as predominantly requiring rote learning and a capacity to write in continuous prose. Barry and Brian disagree and argue that there is a significant decrease in continuous prose responses required in GCSE biology papers, a change that again they reflect in their teaching and assessment. The teachers' views of each other's subject difficulty are likely to emerge in discourse with students. This is particularly likely when they are advising students about appropriate science course options for Years 10 and 11, and when discussing tier entry decisions for Double Award where the awarded grade depends on relative performances in the separate science subjects. The nature and extent of its impact on students' science course choices and teachers' tier entry decisions is beyond the scope of the current research but it is likely that there is an effect there to investigate.

7.4.8 Gender : difficulty interaction

The teachers first raised gender as an issue when discussing their subjects and students' perceptions of their relative 'difficulty'. Teachers do not refer to gender but rather to sex groups i.e. boys and girls as a whole; they tend not to think of gender as a phenomenon that emerges in social interaction. Only Brian amongst the biology teachers refers to boys and girls and that is in the context of his subject gaining more public esteem in recent years because it is perceived to have become more 'difficult', a view supported by my technical findings showing a decrease in mean grade relative to the other sciences from 1995. Brian cites an increase in his male students wishing to take 'A' level biology as evidence of biology now being held in higher esteem, although this might also reflect the shifts in the domain consequent on the introduction of new biology syllabi, which boys might find particularly

engaging or simply to the positive role model that Brian with his PhD and Head of Science Faculty status presents for the boys in his school. In School 3 it appears that the traditional view of biology as a 'girls' subject' has been mediated.

The issue of students' gender is of significantly more concern to the physics teachers than the others during the interviews. Interestingly my technical findings showed that girls significantly and consistently underperformed boys in WJEC GCSE physics and to a less significant and consistent degree in SEG GCSE physics, a finding not shown for biology and chemistry. All three physics teachers talk about gender in relation to students' perceptions of the difficulty of physics, particularly in terms of its mathematical demands, and in relation to attitudinal factors such as confidence and motivation to learn. They all believe girls view physics as being more mathematically challenging than the boys. Unanimously they do not see this being due to girls having less well-developed mathematical skills than boys, and I note that my technical findings did not indicate girls as significantly underperforming boys in their GCSE mathematics grades, although as discussed in Chapter 5, I question the dependability of these findings, but rather to significantly lower levels of confidence in their ability. All the physics teachers believe that girls' lack of confidence results in them tending to 'play safe' and not wishing to be entered for Triple Award Physics, or if they do, they prefer the foundation tier to the higher tier physics papers, nor do they continue with physics at 'A' level to the same extent that boys do. The mathematics : physics interaction and its association with 'difficulty' reinforces the view that physics is for boys.

The three physics teachers mediate these gender effects in similar ways. They all claim to offer more verbal support to girls than boys and try to persuade them of their ability to achieve well in physics. How this influences the physics teachers' tier entry decisions is unclear but it is apparent than girls are more likely than boys to refuse their teachers' wish to enter them for higher tier physics GCSE papers. The physics teachers offered more perspective on girls' attitudes to learning than the teachers of biology or chemistry. They refer to girls being more motivated to learn than the boys, as having better listening skills and in wanting to get things '*right*' in their coursework. However, these attitudinal factors are not cited by the physics teachers as influencing their course or tier entry

decisions or as resulting in significantly different grade outcomes for boys and girls. Cathy (School 1) was the only other teacher to raise girls' relative lack of confidence in their ability. For these reasons Cathy considers differentiation by tiering and a three tier system is more problematic for girls because they lack confidence to enter for the higher tier and tend to play safe with their preferred middle tier entry. Thus for Cathy tiering choices by the students can mediate their achievement; that is their grades may not reflect their achievements because of the ceiling placed on them, raising questions about comparability and 'gradeness'. Cathy, however, is unique amongst all of the interviewed teachers in stating that she takes boys lower motivation into account when making GCSE tier entry decisions: she '*knows*' the boys will make more effort after the mock GCSEs in Year 11 and will enter them for tiers higher than their mock results indicate as being appropriate. Interestingly her views of gender allow her to compensate for boy's lack of motivation but not for girls' lack of confidence as she enters girls for lower tier when they lack the confidence, not ability, to be entered for higher tier. Clive shares a similar view that boys are more likely to underachieve than girls and relates this to their being '*laid back*' but believes '*able boys*' by the time of the examinations '*come up to scratch*'. Whether this influences his tier decisions is unclear. Cathy's comments suggest that gender does mediate her practice and appears to result in her reinforcing the view that girls can't do science but boys can.

Teachers talk of the need for accuracy in their tiering decisions and monitor that their predictions match the outcomes achieved. They seem unaware of the potential for outcomes to mask ceiling and floor effects. Furthermore, while they consider that students' views of themselves influence tier entry decisions they do not reflect on the potential for their decisions to be mediated by their beliefs about for example what a successful physics student is like, or if they associate poor motivation or effort and low confidence with low ability. As noted in Chapter 3 there is evidence from research that tiering decisions are influenced by teachers' beliefs about students that are not to do with their prior achievements and therefore social factors further undermine the meaning of 'gradeness' and grade comparability.

7.5 Reflection

By illuminating the interrelationships between arenas and settings the qualitative study has highlighted the variations that may exist within a population of students taking the same examination, for example variation in their history of access to science courses to meet their developmental needs and in the types of information used to judge their achievements in either science or the separate sciences in their schools and decide their GCSE tier entry. In my technical study I sought to control for variation in my compared populations by only taking those that had sat the same tier of examination papers (Chapters 3 and 4). This aspect of the technical approach to investigating examination comparability, like others discussed in Chapter 3, is challenged by the qualitative study for assuming homogeneity across examination populations drawn as they are from a variety of schools that arguably may reflect some of the same types of arena and setting mediations shown to exist within my three schools. This again challenges the dependability of the outcomes from a technical approach to comparability studies.

CHAPTER 8

Overview and discussion

This final chapter provides:

- an overview of the key findings about the technical and social dimensions of subject comparability within the context of GCSE assessments and offers insights into the social process by which grade outcomes emerge in national assessment at 16+;
- a discussion of the study's limitations;
- recommendations for future practice and research.

8.1 Findings and their implications

Public interest still centres on the key examination comparability issues illuminated by this study's findings as illustrated by the media's annual reporting on GCSE results and teachers' concerns:

> *More than half of all teachers think that exam papers in their subjects are harder than in others ... more than a third said that they were uncertain about the validity of GCSEs.*
>
> Bloom, A. Ipsos Mori Survey, TES, 14 March, 2008, p. 6

> *... new* [GCSE] *guidelines which strip all the mathematical content from some science papers were forcing more schools to use the iGCSE as a real preparation for 'A' levels.*
>
> Daily Express, 3 September 2005, p. 9

Each August, when GCSE results are announced, the media highlights the debate about comparability of standards, schools are compared with one another, students compare themselves with each other and look to their achieved grades as measures of common currency for their future educational opportunities. Examination comparability remains a key concern for teachers, students, parents, employers, educationalists, examining group personnel and educational policy makers.

The initial aim of my research was to use quantitative analysis to illuminate the notion of subject 'gradeness' being stable across subjects. At that point I had an open mind about the goal of comparability. To extend the potential for illumination I controlled for more variables and provided more cross-references to examination paper demands, centre characteristics and coursework factors

in my investigation than other studies of GCSE science subject performance hitherto available in the public domain. I used the common statistical treatments currently in use by GCSE examining groups for investigating comparability in examination performance. I had to obtain information about these treatments from the Director of Research at WJEC as at that time this was not in the public domain; recently, Newton *et al.'s* (2008) book has made this type of information available to a wider audience than GCSE examining group personnel. In addition to those routinely used I also used kappa as it is regarded by some statisticians (Bell, 1999) as the most appropriate statistical treatment for searching for agreement in achieved grades across subjects.

In the beginning I anticipated that my statistical investigation would provide insights about my particular data set that would contribute to the then ongoing technical debate about grade comparability. Because of illness my research was interrupted. This allowed me time to read more widely and to reflect on my quantitative findings, and as a consequence, I came to the realisation that the goal of examination comparability was unachievable. This reflected my emerging understanding that the process by which students gain access to particular grade ranges and their interaction with examination papers are subject to a myriad of social influences and experiences that cannot be addressed by technical treatments however complex. I therefore recast the way I interpreted the quantitative findings. Rather than taking them as the basis for high level inference, which seemed inappropriate given the variety of issues I had found to undermine this, I treated them as offering insights into influences that might disrupt the way that science was represented in examinations over time, within and between science subjects, and students' interactions with these changing representations. In so doing, my interpretation of the quantitative data offered insights into the way that assessment processes play out in teachers' practices, and importantly, insights into the impact that changes in representations of science following changes in the curriculum and examining groups may have for students' achievements.

Comparability is shown to be undermined by technical influences in the quantitative study and by social aspects of the assessment process in the qualitative study. Some of these social aspects are also shown to reinforce the identified technical influences. Jointly and separately the quantitative and qualitative studies illuminate a complexity of inter-relationships that undermine the validity of comparing subject performances and in particular, doing so across time. Despite the

availability of numerous technical treatments it is simply not possible to control for all the varied technical and social influences on examination performances that can affect the validity of comparing examination performance based on the assumption of stability in gradeness across subjects. To say that differences in groups of subjects' grade distributions reflect differences in the subjects' severity of marking, uses the discourse of examining group personnel and either neglects or assumes the constancy of a myriad of potential effects and relationships that are sources of invalidity in relation to comparability of subjects. Furthermore, it does not take account of the social nature of the assessment process, aspects of which emerged in my qualitative study. Jointly my quantitative and qualitative analysis reveals the nature of examination comparability as a chimera.

The quantitative study shows that even taking populations that have been entered for the same tier of papers and comparing their performances across years is undermined by changes in their associated examination centres and the influences that their nature exerts on performance (SRAC, 1976). As in other studies (Smith and Tomlinson, 1989; Nuttall *et al.*, 1989; Drew and Gray, 1990, 1991; Troyna, 1991; Stobart *et al.* (1992); Elwood, 1995; Arnot *et al.*, 1996) the technical analysis shows sex sub-group effects are masked by overall performances. It begs the question what other types of sub-group effects may be masked by just comparing whole populations' performances and adds weight to the thesis that a technical approach and findings are limited in their dependability.

The qualitative findings support the possibility of sub-group effects but in regard not only to students but to their treatment by teachers. For example, Clive talks about the analytical demands of chemistry in comparison to biology and how this influences his preparation of students to understand what GCSE chemistry questions are requiring them to do. The quantitative findings also include girls significantly underperforming boys on WJEC GCSE physics examinations. All three physics teachers reported girls generally lacking in confidence when faced with the mathematical demands of physics GCSE questions. They responded in varying ways, for example Paul offering extra mathematics for physics lessons and being particularly supportive of girls in class, whilst Philip emphasized his strategy of providing a 'secure' and non-threatening environment for his physics lessons and being extra supportive of girls' efforts in class. However, this was not considered to be the case for *all* of the girls they taught. Similarly, not all boys were

reported by the teachers as being able to cope better than the girls with the mathematical demands of GCSE physics examination questions.

The quantitative findings show fluctuations in the science subjects' apparent difficulty across time. This counters the notion of a subject being inherently more difficult than others and also challenges how comparability data may be interpreted across time. Furthermore, differences in views of which science subject is more challenging emerged across the teachers interviewed. These views were variously based on a complex interplay of beliefs about a subject's inherent difficulty, the demands made by syllabi and examination papers, coursework requirements, and students' views of subject difficulty. For example curriculum changes influencing representations of biology, chemistry and physics are shown to undermine the notion of 'gradeness' across time. The thesis includes the first examination of GCSE syllabi (1995) reflecting the introduction of the National Curriculum. Each of the technical treatments showed changes in how the examination performances considered related to each other. For example in the 1995 data set there was a convergence in standard deviation values for the three science subjects and greater similarity in their examination paper skill demands compared with previous years. I argue that the nature of the science subjects was recast consequent on syllabi changes at this time. Certainly this was the view expressed by several of the teachers. Barry (School 1) and Brian (School 3) expressed concerns about WJEC biology examinations becoming cognitively more challenging across time, echoing the anecdotal evidence of other biology teachers referred to in Chapter 1 and reflecting the findings from the technical analysis of the 1993-1995 examination papers. Clive (School 2) also reported that his subject, chemistry, had changed in its nature since the introduction of the National Curriculum, with substantial amounts of chemistry, which he deemed necessary for 'A' level chemistry preparation, being replaced by geology.

Paul (School 1), Peter (School 2) and Phil (School 3) all expressed concern about physics becoming 'easier' in recent years. They noted that the range of concepts included in the syllabi had reduced with 'difficult' concepts such as geometric optics being removed. Peter (School 2) talked at length about the way in which computational demands in questions have been reduced with questions being more structured and less open-ended. Recall of complex physics formulae was also no longer required as these were now provided on the examination papers. In this sense, the

combined quantitative and qualitative findings suggest that physics as represented in examination papers was changing and grade distributions shifting to become 'easier' in relation to biology and chemistry from 1993 to 1995. This interpretation, however, is based on overall performance and says nothing about particular students' interactions with items. From these teachers' views 'gradeness' is seen to be undermined because the grades do not have the same currency as before. The technical study's finding of fluctuations in the science subjects' apparent difficulty across time counters any notion of a subject being inherently more 'difficult' than the others and challenges the ways by which comparability data may be interpreted across years by highlighting the potential for curriculum-examination interactions.

The stability of the correlation findings for the physics and chemistry, and mathematics and physics pairings across time indicates an influence which elicits similar student-examination interactions across these subject pairs. Certainly the teachers tended to report that these paired subjects called on similar skills. Further, this pattern in correlation remained a significant influence on examination performance in the presence of many other influences such as the curriculum changes consequent on the introduction of the National Curriculum. If GCSE subjects vary in their skill requirements, the grades can only still claim to have common currency across these subjects if the grade awarding process facilitates this as claimed by examining groups. As discussed in Chapters 2 and 3, examiners function as a 'guild of professionals' (Sadler, 1987) with an understanding of their particular subject and how examination performance equates to specific grades. However, they do not have such an understanding for other subjects. The dependability that is assumed in 'gradeness' is that separately professional judgement can decide what performance reflects a particular grade, and that this professional judgement is consistent across subjects and moderated statistically with regard to relative performance over time. Arguably, however, a grade A in physics cannot have the same currency as a grade A in biology when they are associated with different skill demands.

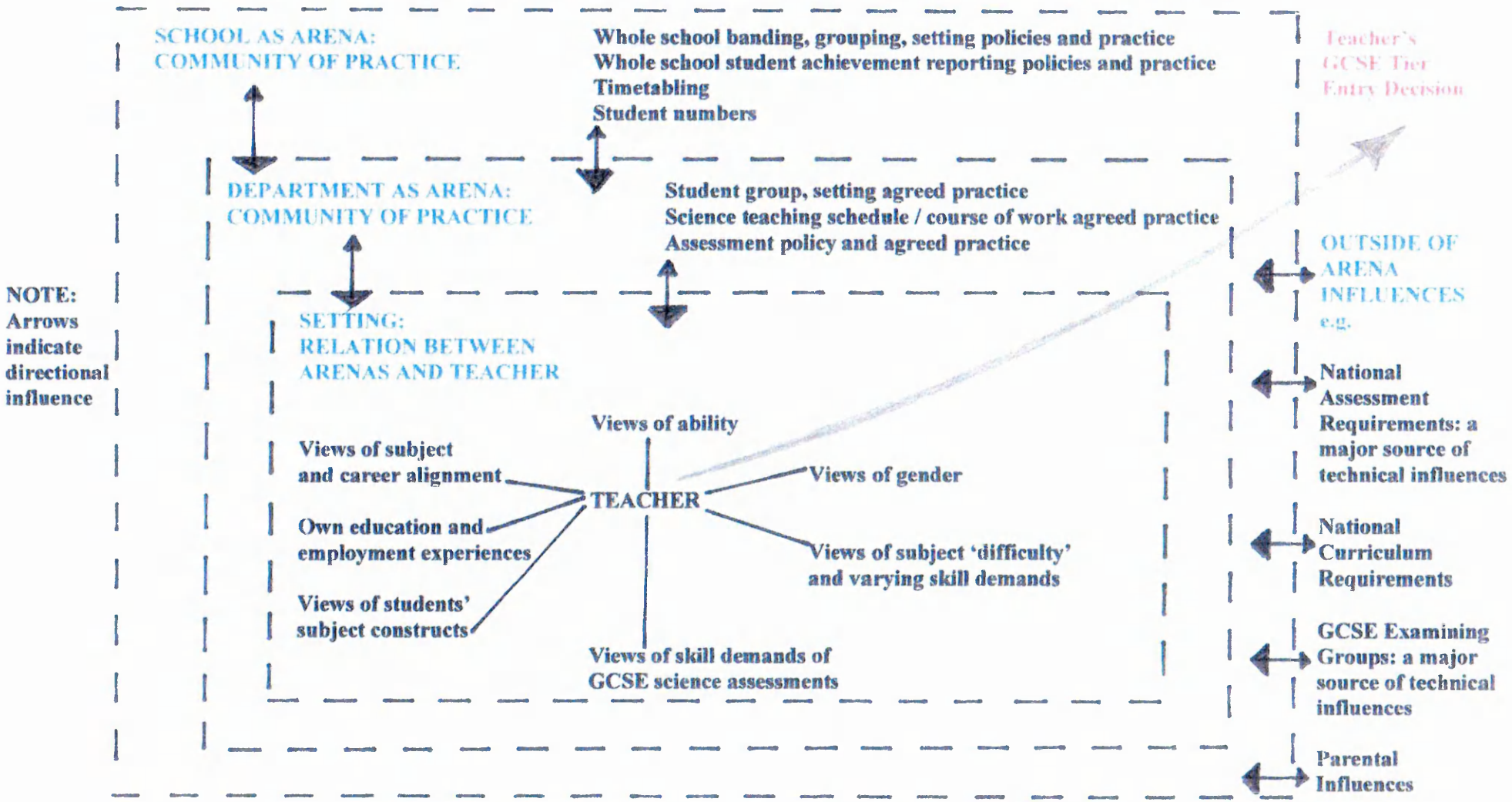
The notion that if a student obtains a particular grade in one science subject it is predictive of the same grade in another, is not strongly supported by the findings across both WJEC and SEG. These findings indicate that students' interactions with examination artefacts i.e. item factors, vary from student to student and population to population. One would not expect identical grades to be

obtained by the same population across different subjects even if these have similar skill demands as there is still the potential for differences in interaction with items, for example in the task that the students perceive and therefore what skills and understanding they use, and / or the contexts used. Girls' lower performance in physics relative to boys despite an overall better performance in all of their GCSE subjects held for all WJEC populations but not for all of the SEG populations. This finding suggests that interactions between students and the specific examination paper varies, and either that girls taking SEG are advantaged in comparison to girls taking WJEC or that boys taking SEG were disadvantaged compared with the boys taking WJEC. The important point here is that 'gradeness' is an assessment artefact.

The qualitative findings discussed thus far offered support for the technical findings but in addition provided further insights about the assessment process in relation to comparability and teachers' relationship with this in the GCSE assessment process. For example, the judgements made by teachers when entering their students for tiers of GCSE science examinations do not take place in a social vacuum. They are shown as being constituted through interaction and as products of teachers' views and beliefs of mind, subjects and gender amongst others, departmental and school practices, and outside of school influences. This study like others suggests that school arena practices influence teachers' judgements about students' entry to GCSE science examinations. However, in addition it reveals something of the complexity of teachers' thinking and practice that culminate in these judgements. The findings indicate that factors that are known to influence students' achievements, for example which school a student attends (Goldstein *et al.*, 1993; Sammons *et al.*, 1995), which set a student is placed in (Ireson and Hallam, 2001; Wiliam and Bartholomew, 2004) may mask more than they reveal.

As the study has progressed, my initial focus on teachers' tier entry decisions extended to reveal a rich picture of setting, arena and social order (Lave 1988) interactions that is largely coloured by constraints and 'ceilings' on students' opportunities to learn. Teachers were shown to orchestrate their settings within arenas (school and department) using assessment and in turn being constrained by assessment in ways that varied depending on the school and the individual. Figure 8.1 provides an overview of the findings from both the quantitative and qualitative analyses of the influences on teachers' orchestration. The thesis has explored the cultural nature of assessment at

FIGURE 8.1 A MODEL OF TEACHERS' ORCHESTRATION OF ASSESSMENT PRACTICE



This model acknowledges but does not represent the influence pedagogic knowledge / school knowledge / subject knowledge of Banks *et al.* (1999) on the setting as beyond the scope of this research.

the level of the practitioner (science teacher) and in doing so, identified the impact that the interactions between arena and setting effects have on students' opportunities to learn and progress. In Figure 8.1 the dotted lines represent the 'permeability' of the social order, arenas and setting to influences that may flow from any one of these to others. The outer level, the social order level, is where social structures such as national assessment requirements and GCSE examining groups' practices exert their influence on institutions. The next box represents the institutional arena and the practices and policies such as school banding and timetabling that are mediated by external influences and in turn mediate the next layer which depicts within the institution the arena of the department. The internal box represents ongoing activity as teachers orchestrate their settings with their students. The arrows illustrate the directional flow of influences based on my findings. For example, student numbers, timetabling, whole school student achievement, reporting policies and practice, and whole school banding, grouping, setting policies and practice are shown to be mainly constraining influences on departmental practice and this in turn on teachers' activity.

The picture that emerges is multifaceted. It reveals teachers' distinctive individual thinking and practice: they have their own concerns about assessment. Teachers construe their subject in their own ways and similarly formulate their thinking in relation to assessment as they choose. They are shown to hold views of ability which influence their practice. For example Clive's belief in fixed ability informs his practice of using cognitive ability tests to assess students and to allocate learning opportunities; this is shown to become the practice of the whole science department in School 2 because Clive uses his position as Head of Science Department to make this departmental policy. Here the influence is from Clive's setting to departmental arena. In School 3 the Head of Science Department, Brian, holds a different view of ability and resists categorizing students, teaching them in mixed ability groups to keep learning pathways and opportunities for progression as open as possible within the constraints of the school arena influences of timetabling and student numbers. In contrast to Clive, Brian's views do not result in him insisting that his departmental members also teach in his preferred mixed ability groups – there is no overt influential flow from his setting to departmental arena in this particular respect. In School 3, the school arena policy of allowing teachers to teach in their preferred grouping arrangements (mixed ability or set groups) is seen to be the major influence on teachers' practice in relation to students grouping in Key Stage 3.

Other influences on teachers' practice within their settings revealed by the qualitative study included their views of: gender; a subject's 'difficulty' in terms of inherent skill demands as well as the skill demands of Key Stage and GCSE science assessments; students' perceptions of their subject; and their own education and employment experiences in relation to their subject. The cognitive skill demands of GCSE examinations shown by the quantitative study to vary in weighting across the science subjects and across time are shown to be an issue that influences teachers' practice. Clive talks about the analytical demands of chemistry in comparison to biology and how this influences his preparation of students to understand what GCSE chemistry questions are requiring them to do. So although Figure 8.1 serves to illustrate the main findings from the qualitative study it also encompasses the illumination of comparability resulting from the quantitative study.

The technical issues cannot be represented as discreet influences, they are best thought of as being embedded as influences throughout Figure 8.1. For example, examining groups play a major role in enacting the national assessment requirements and this is why they are positioned in the social order in Figure 8.1. The research has shown that national assessment requirements are mediated by examining groups. For example in Chapter 4 differences in how WJEC and SEG differentiate their science examinations were identified in the quantitative study. *How* the examining groups administer the national requirements is shown to be important to teachers in the qualitative study with teachers choosing an examining group according to several issues, for example their views of learners, the cognitive skill demands of the examination papers, the appropriateness of the moderation of coursework, access to examining group offices, the provision of a particular assessment approach (linear or modular), the timely availability of syllabuses and support materials, and familiarity with examining group personnel are all highly valued. In this way teachers mediate the assessment process from their setting out to the social order, from the 'inside to the outside' in Figure 8.1. Phil's actions in raising the disproportionately low percentage of Grade As in WJEC GCSE physics compared with that in other GCSE examining groups was followed by WJEC altering their practices to bring their percentages into line with those of other groups. Teachers' thinking and practice are also shown to be changed by local and national issues, for example by following a tradition of using students' Single Award results to inform tier entry

decisions for Triple Award in School 1 and selecting only biology and physics coursework for Double Award moderation in response to *'harsh'* chemistry moderation.

A key finding too was that teachers' assessment practice was most commonly constrained by the arena practice of timetabling. The findings revealed that KS2 SAT results informed decisions in the late 1990s and early 2000s about the allocation of students to groups as they entered secondary school. The interactions between these groupings and KS3 SAT tier allocation effectively 'locked' students on to assessment and curriculum pathways from Year 7, pathways which school structures, for example banding and the curriculum's timetable, made it almost impossible to break away from. Teachers responded differently to this and the consequences for students' access to learning opportunities varied. Whatever the response to this constraint, with the exception of Brian and to a lesser extent Barry, teachers' discourse and practice were dominated by reference to SAT levels and testing. Teachers were preoccupied with categorizing their students *as* SAT levels, these levels being agreed upon in their meaning and value for comparing students and determining their potential to learn. Foucault's comment appears apt here:

> *The school became a sort of apparatus of uninterrupted examination ... It became increasingly a perpetual comparison of each and all (pupils) that made it possible both to measure and to judge ... a constantly repeated ritual of power*
>
> (Foucault 1977: 186, 192)

The qualitative findings provide support for Torrance's (2007) argument, based on his research, that the practice of assessment has moved from assessment of learning, through assessment for learning, to assessment as learning, with assessment procedures and practices coming completely to dominate the learning experience and 'criteria compliance', which may be regarded as the statements of attainment within SAT levels in this study, replacing 'learning'.

What appears to be a preoccupation with SAT levels rather than students' developing understanding of particular scientific concepts is a preoccupation with obtaining information that enables teachers to *compare* students. Such practices result in students' allocated SAT levels largely determining their teaching group placements, which in turn determine their opportunities for learning and eventually, access to GCSE assessments offering specific tiers of awarded grades. This pre-occupation with comparability stems from the need demanded by social order level

policies to justify to themselves, their school managers, parents and educational and political bodies, their decisions for allocating learning opportunities within limited resources. The teachers in the main preferred to prepare students in any group for one level of SAT papers or tier of GCSE. This practice is largely viewed as an efficiency model by the majority of teachers within the study because only work associated with one particular tier needs to be taught and differentiation, which is viewed as problematic, is kept to a minimum. Getting the students' group allocation 'correct' in the sense that the teachers' view of a student's capacity to learn the work associated with a particular tier matches that of the group, was essential in teachers' minds. The study's findings show that if students are deemed level 5 at the end of Year 9, then that extrapolates to a particular GCSE tier for Year 11, and that is what they are entered for.

> *At present, schools have to make difficult decisions on the selection of pupils* [students] *for GCE or CSE courses, perhaps as early as the end of the third year ..... Early choices cannot allow for the development of pupils' abilities ..... many teachers would say that the task ..... is one which they find particularly difficult and unrewarding ..... this practice may, to some extent, pre-empt the results of the examinations themselves.*
>
> (Schools Council, 1975, p. 9)

The above comment appears to apply as much to today in relation to tiers of GCSE examination papers, as it did in 1975 for 'O' level and CSE.

8.2 Limitations of the research

The limitations of the research can on one level be interpreted in terms of the number of sources of data, both in the quantitative and qualitative studies. In terms of limited data access from a technical standpoint, more years' worth of data, for example including quantitative examination performance data up to and including the years when I interviewed teachers in my qualitative study, would have been useful in identifying the existence of any continuation, and hence significance, of the trends in examination performance revealed by my technical analysis of the 1993 – 1995 data. For similar reasons, the analysis of examination performance data from additional GCSE examining groups would have enabled the similarities and differences in the technical findings from my comparison of WJEC and SEG data to be enriched. For example, there

were significant differences between WJEC and SEG in the relative performances of boys and girls in GCSE physics examinations. Girls did better in relation to boys on SEG but not on WJEC GCSE physics examinations, arguably indicating that something within the WJEC assessment process, for example the physics examination questions, consistently disadvantaged girls or advantaged boys across the years of my data collection. For reasons of time, resources and access to data by all other existing GCSE examining groups, my quantitative study was limited to 1993-1995 and WJEC and SEG data. As it was I had an enormous amount of quantitative data to handle, most of it received in paper format and not on computer discs. It is doubtful whether I could have managed more data in the time available to me. The particular data used could be seen as having little significance at this point in time. However, as it coincided with significant major national curriculum and assessment changes it is of particular value in illuminating sources that undermine the stability of grades over time within science subjects.

The limitations of the research can, on another level, also be interpreted in terms of limitations of the statistical techniques themselves, which have been discussed in Chapters 3 and 5. Given the sociocultural view that meanings are negotiated and therefore that constructs emerge rather than are given, there is the potential for differential individual student interaction with assessment items. I, therefore, view the evidence of sex sub-group effects as limited in their ability to reveal gender effects as the groups were treated necessarily as homogeneously because access to individual data was not considered with reference to individual interactions.

In terms of limited data access from a qualitative standpoint, in Chapter 5's discussion of my sociocultural position and meaning of case study, I state that there was no intention of defining the cases as what they might be 'cases of' other than as individual science teachers and my findings from the qualitative study are not assumed to be generalisations – they do not necessarily apply to *all* science teachers. I have looked across the teachers' accounts for enduring practices and shared beliefs and presented these in Chapter 7 but, as discussed in Chapter 5.2, social phenomena are neither time nor context free and generalisations are impossible. Whether my findings from interviewing my nine science teachers are *transferable* is for others to decide. My cases are not representative but neither are they in any way atypical. A sociocultural approach anticipates commonalities in the social mediation of structures such as national assessment and examination

policy and practice but also anticipates that these can be disrupted at arena level and by individuals in settings as my data established. The findings therefore are about the impact of these social order and institutional influences on teachers' practice and learners' opportunities to learn and how this mediates their achievements. In this sense there are messages for all schools where such structures apply. Interviewing more science teachers than the nine included in this thesis would have provided more personal accounts of how teachers respond to arena level structures and practices and the social order as they orchestrate their settings, and increased the potential to reveal more about the complexity of interactions shown by this thesis to constitute that orchestration. As it was, time and resources limited my interviewing to nine teachers. Furthermore, for reasons of limited time and resources I have gone to teachers and not to students to explore the sociocultural nature of assessment and its impact on comparability.

8.3 Recommendations

As discussed in Chapter 3, the decrease in examination comparability research in the 1980s was largely due to criticism regarding the limitations of the available statistical treatments for producing what was then regarded as valid outcomes. Efforts have focused on ways of taking account of social influences within the statistical treatments by using complex mathematical treatments as in multilevel modelling (Goldstein, 1995) or by obtaining more and more information about the candidatures for interpreting and qualifying GCSE performances during and after grade awarding. As the co-ordinator of the Research and Evaluation Division of University of Cambridge Local Examinations Syndicate notes:

> ... *the statistical information available at grade awarding has become increasingly more sophisticated and this has an impact on comparability because awarders are more aware of how the quality of the entry* [candidature] *for a specific specification* [syllabus] *differs from that of the whole entry.*
>
> Bell, 2005

However, the *'quality'* referred to above still appears to relate to those influences that can be reduced to technical treatments:

> *The quality of the entry is based on measures of prior attainment, i.e. KS3 at GCSE and GCSE at 'A' level. These are used to generate putative grade distributions. ... The awarders can compare this putative grade distribution with the actual results from previous years (to see if the entry has changed) or with the predictions for all awarding*

bodies (to compare with other specifications). There is a problem with KS3 because independent schools do not necessarily take the test. This complicates the interpretation of the figures.

Bell, 2005

I have argued that GCSE statistical comparability of performance studies cannot achieve enhanced validity for examinations by controlling for factors such as question ambiguity, gender and cultural bias, and examination cognitive demand. Even if it were possible to do so, it would still leave the fundamental flaw in GCSE examining, that grades are assumed to have stability and meaning across students, subjects and time, and that performance outcomes may be treated as independent measures. This indicates a continuing reliance on the psychometric tradition and the view of mind as fixed with innate predictive 'intelligence' underlying it. This is despite a move towards educational assessment as evidenced for example by the incorporation of coursework.

Assessment [within the UK] *is undergoing a paradigm shift, from psychometrics to a broader model of assessment, from a testing and examination culture to an assessment culture.*

(Gipps, 1994, p.1)

The paradigm shift referred to by Gipps involves more than innovatory modes of assessment. Rather it embodies new conceptualizations of learning and achievement and their assessment. In turn, these may be viewed (Broadfoot, 1996) as a reconceptualization of the purposes of education and its mode of delivery in response to society's changing industrial culture. Thus conceptually and practically, education and assessment may be regarded as being inextricably involved with social, economic and political factors (ibid.). As Sutherland (1996) writes:

The practice of assessment is "socially embedded": people do it to other people – or at least to other people's children. We shall only understand it fully if we take account of the social, economic and political contexts in which it grows and is practiced.

(Sutherland, 1996, p. 19)

My view is that the paradigm shift referred to by Gipps is ongoing and remains so in 2008. Such shifts, I argue, are characterized by tensions emanating from a complex interplay not only of social, economic and political influences, but also from a multitude of technical assessment issues. In common with Firestone (1998) I perceive these influences and issues as being associated with

particular groups of people such as politicians, educationists, teachers and assessment technicians. These different groups:

- function in different arenas;
- align themselves with different curricular and assessment issues;
- use language which differs in its nature and emphasis;
- measure success by different means;
- differ in the degree to which they are willing and/or are able to interact with one another.

Such group differences, I argue, are potential sources of tension. Different groups have different concerns and different ways of dealing with them. The groups differ in size and in their capacity to organize themselves to operationalize their aims and objectives and communicate these to other groups (Firestone, 1989). At any point in time one or more of these groups may dominate in determining the separate or collective development of assessment ideology, policy and practice and therein, lays the potential conflict. The exploration of group differences, emergent tensions and conflicts permeates this thesis.

Educators and assessment technicians will have different priorities related to the functions of their roles. Assessment technicians function in different arenas. Assessment technicians might perceive the creation of differentiated examination papers in science as a solution for the technical problems associated with assessing a student population of diverse numerical ability. However, this 'solution' might generate ideological and practical challenges for teachers. Continuing with this example, when teachers have to allocate each of their students to particular tiers of examination papers, what are the consequences for each student's future learning opportunities? How does this impact on teachers' aspirations for their students? On a practical level, how are teachers to manage the teaching of students who need to be entered for different examinations? How does this management impact on the social dynamics within a class of students? Furthermore, teachers' perceptions of subject difficulty may be influenced by the nature of examination papers constructed by assessment technicians, with a consequent impact on students' access to learning. For example, a particular subject might be associated with examination papers relying on structured questions; another subject's examinations might place a greater emphasis on extended free response questions. Teachers might view the former subject's examinations as being 'easier' for students

with poor literacy skills than the latter. In such circumstances teachers may advise students about subject 'choices' they believe will maximize their grade achievements and in that way mediate students' access to learning and subsequent achievements. As Firestone (1998) writes,

> *This overlay of different perspectives compounds the difficulty of resolving purely technical* [assessment] *problems.*
>
> (Firestone, 1989, p.188)

The recent publication, 'Techniques for monitoring the comparability of examination standards' (Newton *et al.*, 2008), is written largely by persons either currently working or with past experience of working within GCSE examining groups. Newton *et al. 's* (ibid.) book is premised on statistical techniques which are still from a psychological tradition with the assumption that a multilevel approach takes account of social, cultural, historical and institutional and individual influences, which it can not. Whilst Newton *et al.* recognize the limitations on these techniques they nevertheless discuss the 'best' approaches to comparability as a political and social necessity.

GCSE examination grades are socially constructed and shown by this research to reflect a complexity of technical and sociocultural interactions. Comparing performances of one examination with another, even from the same subject domain, does not necessarily reveal differences in their difficulty. From my sociocultural perspective, and as this thesis has shown in both the quantitative and qualitative studies, first, comparability is a chimera, and second, and importantly, treating students and their assessments as being comparable gives them undue status in teachers' eyes and practices and this impacts on students' access to learning opportunities. The detrimental impact of this belief in comparability of students and in the meaningfulness of grades and levels from this thesis's point of view challenges the search for more and more technical techniques without recognition of the limitations of their outcomes as measures of what students can know. I argue that they are not even measures of what they *do* know. In Newton *et al. 's* book it is clear that the political agenda is prioritized by examining groups, as it has to be. However, this promulgates for teachers and other users of assessment, assessment measures as a way of judging human achievement and potential.

All of this leads me to recommend teachers, parents and other users of examinations that they interpret examination pass rates with care and to be cautious in the actions they take on the basis of

them. It also leads me to recommend that examining groups think about ways of presenting assessment data that are more socially just so that their consequential validity is enhanced. The current popular way of presenting percentages of A* – C achieved grades in different subjects and in different schools for comparability purposes does not take account of disparities in entry rates, different types of sub-group effects, and curriculum-assessment interactions. Sub-group disparities in examination performance, particularly gender disparities in physics, need to be monitored effectively by examining groups. This recommendation for examining groups is particularly important as routes to gaining science qualifications at 16 plus increase in variety, for example with the current rolling out of the new 21st century science examinations.

I was privileged to be given access to examination performance data by the examining groups WJEC and SEG. Other GCSE examining groups that I approached at the time of my data collection refused access to this type of data and gave no rationale for doing so. Confidentiality of students' and examination centre identities is an important issue when giving access to this sensitive type of data. Furthermore, there is the ethical issue of students and centres providing *their* agreement to the release of their examination performance data. If examining groups were to extend researchers' access to assessment outcome data it would broaden and enrich the research conducted within the field of examination performance, and assessment in general, by bringing into the field researchers outside of GCSE examining group employment who possess a variety of theoretical positions. Given the commercially competitive nature of GCSE examining groups, my recommendation for greater access to assessment outcome data is, however, unlikely to be fulfilled.

A lack of access to information about the techniques used by examining groups to research examination comparability was a problem for the research, as discussed in Chapters 3 and 4. Examination performance research is nearly always carried out by current or past examining group employees, and without the Director of Research at WJEC providing me with information about the Group's analytical techniques, this thesis would not have been possible. There is little public discussion or knowledge of the technical practices of examining groups in relation to grade awarding and their activity for checking on comparability of grading severity within and between subjects across time. This echoes my earlier discussion about groups of people functioning in specific arenas, using language that differs in its nature and emphasis from that of other groups and

differing in the degree to which they are willing to interact with other groups, for example with teachers and educational researchers. GCSE examining groups are under intense pressure from the media regarding their practices. One might speculate that under this intense pressure they seek to retain their 'power' by keeping knowledge of their practices to themselves, as was my experience as a researcher. However, their practices are also mediated by wider social influences such as the public and political response to their outcomes. Therefore, they need to maintain user confidence in the validity of their practices and their outcomes. This situation leads me to recommend that GCSE examining groups make information about their analytical techniques, and in particular about their comparability processes, more accessible to those outside of their employment. Commendably, the recent publication by Newton *et al.* (2008) goes some way towards fulfilling this recommendation, although awareness of and access to the publication itself needs to be improved.

In the early 1990s it was possible for schools to 'shop around' for their GCSE science syllabuses and this is still the case in 2008. Schools would enter some students for Triple Award with one particular examining group and other students for Double Award science with another. This phenomenon was prompted by a belief that it was easier to obtain a high grade in some syllabuses than in others (OUDLE, 1995; SEG, 1996; WJEC, 1994). From 1995 students presenting themselves for the separate sciences at GCSE were required to take all three with the same examining group. Nevertheless, this phenomenon of schools 'picking and mixing' GCSE science syllabuses from those available by the different examining groups still occurred and triggered a rise in interest about comparability of grading standards by the School Curriculum and Assessment Authority. Inter-group comparability exercises (SEG, 1995) were instigated in response to concerns about this issue. These have significantly decreased since 2000 resulting in less information about examining group practices being in the public domain. A possible reason for this is the sub-committee that organised such activity was disbanded because of the re-organisation of the Joint Council to form the Joint Council for (General) Qualifications. Although the sub-committee has been reformed, it focuses on screening examinations for comparability using statistical analysis. The previous types of cross-moderation exercises are now infrequently conducted and only instigated if the statistical findings indicate significant disparity in grading

(Bell, 2005); the reason provided is inter-group cross- moderation is too costly. This practice has reduced examining groups' consideration of the sociocultural dimension of GCSE assessment, as only when cross-moderators come together can there be a discourse that encompasses this dimension in relation to the considered examination papers and awarded grades. If one adopts the notion of grade awarders as a *'guild of professionals'* (Sadler, 1985) who are guardians of 'standards' of grading across time, the reduction in inter-group cross-moderation appears to be counterproductive for answering the ever increasing voice that 'standards' at GCSE are falling.

Newton *et al* (ibid.) discuss using techniques to make examination comparability valid. I have taken a different position in this thesis: I have critiqued the act of comparing examination performances. For me examination comparability is not possible and there is a need to understand 'gradeness'. I have used my technical analysis to explore *in*comparability and in so doing, raised issues that influence 'gradeness' and, in addition to that from my theoretical position, shown how teachers and students influence comparability. Consequently, on the basis of the thesis findings about the social nature of assessment practices and judgements it would benefit the field if examining groups and their regulatory body instigate more activities such as inter-group cross moderation that engage examiners in discourse regarding the social mediation of GCSE assessment and the consequences for the meaning of gradeness.

My study only highlights parental influences on assessment within settings and arenas in respect of teachers' responses to the pressures they exert on teachers' group and tier entry decisions for their students. It is reasonable to assume that the shaping of students' views of teachers' assessment practices draws on influences that include those from their families, friends and wider communities. There is scope for further research on how influences that include those from students' families, friends and wider communities shape students' views of teachers' assessment practices and their motivation to study science - and provide opportunities for identifying how parents might best be provided with and respond to schools' formal systems of reporting achievement in secondary schools.

In particular, this study had identified the scope for substantial further research within arenas and settings on understanding the nature of these relationships at practitioner and student levels and their impact on students' opportunities to progress. Given my limited resources I decided to focus

on accessing teachers' thinking and practices, not those of students for investigating how assessment plays out in schools and the influences on students' access to and performance on GCSE science subjects. Studies, for example Elwood and Comber, 1996, Boaler *et al.*, 2000 and Boaler, 2008 have gone some way towards exploring students' perceptions of teachers' assessment. Substantial research opportunities remain for improving an understanding of students' perceptions of assessment practices and how these influence students' learning and achievements. More research within settings and arenas that follows secondary school students' assessment and learning pathways is needed particularly as routes to 16 plus science examinations increase in variety. It is less clear what influence examining group and syllabus choice have on students' achievements. One might speculate that there is an effect there to investigate not least in terms of the influence a modular rather than a linear approach might have on a students' capacity to recall knowledge during examination.

The study revealed that several of the teachers' based their practice on assumptions of fixed potential. In contrast, one teacher in particular, Brian, based his practice on the premise that students' achievements were a matter for teaching and did not place ceilings on them. Hence he resisted their early categorisation and the subsequent constraints on their learning opportunities by opting to teach them in mixed ability groups in the early years of secondary schooling. Figure 8.1 reveals this effect in the reverse directionality of the arrows. Brian was unique amongst the interviewed teachers in taking this stance about the relative importance of teaching. His comments show that teachers are engaged in thinking about the effect that choice of examining group may have on their students' GCSE performances and so their concern for examination comparability. Based on my research it would be helpful if teachers were provided with the time and support to reflect on their educational beliefs and practices and how these influence their students' learning and progression as part of a programme of continuing professional development. This support should come from outside the school in which they teach to avoid promulgating the status quo and to encourage personal review in a non-threatening environment.

The relationship between teachers' assessment-related activities, school practices, and influences from beyond the school is recommended as being ripe for further research. In particular there is a need for a greater understanding of the importance of school structures and practices on

influencing students' opportunities to progress. For example, the introduction of German in School 1 appears to have a more significant constraining effect on students' opportunities to learn science than the structures used to introduce a second foreign language in Schools 2 and 3. How extensive is this constraining effect and how could it be ameliorated? I would say to teachers, and in particular to head teachers responsible for school policies, they need to examine their arrangements for grouping students on entry to and throughout secondary schooling to identify relationships between flexibility of students' group movements, development and equity of access to curriculum provision.

The system of statutory national key stage tests in Wales was, until 2000, the same as in England. In 2000 The National Assembly for Wales took responsibility for these tests at which point they were developed by test agencies on behalf of ACCAC, whilst those in England were developed for the QCA. Following the outcomes of the Daugherty Report in 2004 commissioned by the Welsh Assembly, KS2 assessments were made optional in 2005 as were KS3 tests in 2006. A new system of assessment for key stages 1-3 is currently (2008) being piloted. These changes beg the questions, what do schools use now to allocate students to groups when entering secondary school, and whatever is used, does it have the same effect as that identified in this study for locking students onto assessment and curriculum pathways? My recommendation is that research into these issues is timely and necessary for informing educational policy makers to avoid placing constraints on students' access to learning opportunities.

And to make an end is to make a beginning.

The end is where we start from.

Little Gidding, T S Eliot

BIBLIOGRAPHY

ABOUSERIE, R. (1992) *Statistics for educational researchers: course guide for M.Ed students.* Cardiff, University of Wales.

ACKER, S. (1999) *The realities of teachers' work: never a dull moment.* London, Cassell.

APPLE, M.W. (1978) Ideology and Educational Reform. *Comparative Education Review*, 26, pp. 367-387.

APU (1985a) *Science in Schools Ages 13 and 15: Report 3 (APU).* London, DES.

APU (1985b) *Practical Testing at Ages 11, 13 and 15: Science Report for Teachers 6 (APU).* London, DES.

ARNOT, M., DAVID, M. and WEINER, G. (1996) *Educational Reform and Gender Equality in Schools.* Manchester, Equal Opportunities Commission.

ASSOCIATED EXAMINING BOARD (AEB) (1985) *Statistics Summer 1994.* Guildford, AEB.

ASSOCIATED EXAMINING BOARD (AEB) (1995) *Personal communication with Director of Research.* Guildford, AEB.

ASSESSMENT SUBJECT GROUP (2003) *Assessment of chemistry in schools.* London, Royal Society of Chemistry.

BAKER, E and O'NEIL, H. (1994) Performance Assessment and Equity: a view from the USA. *Assessment in Education*, 1, pp11-26.

BALL, S. J. and GOODSON, I.F. (1985) *Teachers' lives and careers.* Lewes, Falmer Press.

BARDELL, G.S., FORREST, G.M. and SHOESMITH, D.J. (1978) *Comparability in GCE: a review of the boards' studies, 1964-1977.* Manchester, Joint Matriculation Board.

BASZANGER, I. and DODIER, N. (1997) Ethnography: relating part of the whole. In D. Silverman(Ed.) *Qualitative Research: Theory, Method and Practice.* London, Sage.

BAUERSFIELD, H. (1988) Interaction, construction, and knowledge: Alternative perspectives for mathematics education. In T. Cooney and D. Grouws (Eds.) *Effective mathematics teaching* (pp. 27-46). Reston, VA, National Council of Teachers of Mathematics and Lawrence Erlbaum Associates.

BEARDSLEY, M.C. (1981) *Aesthetics: Problems in the Philosophy of Criticism.* Indianapolis, Hackett.

BEIJAARD, D. (1995) Teachers' prior experiences and actual perceptions of professional identity. *Teachers and Teaching*, 1(2), pp. 281-294.

BELL, J.F. (2005) *Personal communication.*

BELL, J.F., (1989) A Comparison of Science Performance and Uptake by Fifteen-year-old Boys and Girls in Co-educational and Single-sex schools – APU survey findings. *Educational Studies*, Vol.15, No. 2.

BELL, J.F., (1999) *Investigating gender differences in the science performances of sixteen-year-old pupils.* Paper presented at BERA, University of Sussex, Brighton, 2-5 September 1999.

BELL, J.F. (1999) *Personal communication.*

BELOE REPORT (1960) *Secondary Schools Examinations other then GCE.* London, HMSO.

BENSON, A. (1993) Unpublished M.Ed. dissertation. Cardiff, University College of Cardiff.

BENSON, A. (1995) *The cognitive skill demands of WJEC GCSE Science Examination Papers.* Oxford, University of Oxford.

BENSON, A. and DOHERTY, A. (1999) Training and Recruitment of Chemistry Teachers. London, Royal Society of Chemistry Chemical Education Research Group (CERG).

BENSON, A., ELWOOD, J. and MURPHY, P. (2005) Assessment Practices in Science: Barriers to Access and Achievement. Paper presented at *European Conference on Educational Research* (ECER), Dublin, September 2005.

BENTLEY, T. (1998) *Learning and beyond the classroom.* London, Routledge.

BILLINGTON, R. (1988) *Living Philosophy: An Introduction to Moral Thought.* London, Routledge.

BLOOM, A. Ipsos Mori Survey, *Times Educational Supplement* (TES), 14 March, 2008, p. 6.

BLOOM, B. S. (1976) *Human Characteristics and School Learning.* New York, McGraw-Hill.

BOALER, J. (1997) *Experiencing School Mathematics: Teaching Styles, Sex and Setting.* Buckingham, Open University Press.

BOALER, J., WILIAM, D. and BROWN, M. (2000) Students' Experiences of Ability Grouping – disaffection, polarization and the construction of failure. *British Educational Research Journal,* Vol. 26, No. 5, pp. 631-648.

BOALER, J. (2002) *Experiencing school mathematics: traditional and reform approaches to teaching and their impact on student learning.* Mahwah, NJ, Lawrence Erlbaum.

BOALER, J. (2008) Promoting 'relational equity' and high mathematics achievement through an innovative mixed-ability approach. *British Educational Research Journal,* Vol.34, No. 2, pp. 167-194.

BOYLE, B. and CHRISTIE, T. (1996) *Issues in Setting Standards: Establishing Comparabilities.* London, Falmer Press.

BREDO, E. (1999) Reconstructing Educational Psychology, in MURPHY, P. (Ed.) *Learners, Learning and Assessment.* London, Paul Chapman Publishing.

BRIGHT, M. (1998) The trouble with boys. *The Observer,* 4/1/98, p. 13.

BRIMER, A., MADAUS, G. F., CHAPMAN, B., KELLAGHAN, T. and WOOD, R. (1978) *Sources of Difference in School Achievement.* Windsor, NFER Publishing Company.

BROADFOOT, P. M. (1979) *Assessment, Schools and Society.* London, Methuen.

BROADFOOT, P. M. (1984) *Selection Certification and Control, Social Issues in Educational Assessment.* Lewes, Falmer Press.

BROADFOOT, P. M. (1994) Performance assessment in perspective, in H. Torrance (Ed) *Evaluating Authentic Assessment.* Buckingham, Open University Press.

BROADFOOT, P. M. (1996) *Education, Assessment and Society: A Sociological Analysis.* Buckingham, Open University Press.

BROADFOOT, P., MURPHY, R. and TORRANCE, H. (1991) (Eds) *Changing Educational Assessment: International Perspectives and Trend.* London, Routledge.

BRUNER, J. S. (1972) *The Relevance of Education.* London, Staples Press.

BRUNER, J. S (1996) *The Culture of Education.* Cambridge, MA, Harvard Univ. Press.

BURGESS, R.G. (1982) *Field Research: A Sourcebook and Field Manual.* London, Allen and Unwin.

BURKS, B. S. (1928) The relative influences of nature and nurture upon mental development. *27th Yearbook National Society Studies in Education*, Part 1, 1928, 219-316.

BURT, C. (1921) *Mental and Scholastic Tests.* London, Staples Press.

BUTTERFIELD, S. (1990) The development of secondary assessment and examinations. In Riding, R. and Butterfield, S. (Eds.) *Assessment and Examination in the Secondary School.* London, Routledge.

CHRISTIE, T. and FORREST, G. M. (1981) *Defining Public Examination Standards.* Schools Council Research Studies. London, Macmillan Educational.

COBB, P. (1999) Where is the mind? In P. Murphy (Ed.) *Learners, Learning and Assessment.* London, Paul Chapman Publishing.

COBB, P. and BAUERSFIELD, H. (1995) The coordination of psychological and sociological perspectives in mathematics education. In P.Cobb and H. Bauersfield (Eds.) *Emergence of mathematical meaning: Interaction in classroom cultures.* Hillsdale, NJ, Lawrence Erlbaum Associates

COBB, P., McCLAIN, K., de SILVA LAMBERG, T. and DEAN C. (2003) Situated Teachers' Instructional Practices in the Institutional Setting of the School and District. *Educational Researcher*, Vol. 32, No.6, pp. 13-24.

COHEN, L. and MANION, L. (1991) *Research Methods in Education.* London, Routledge.

COLE, M. (1996) *Cultural Psychology.* Cambridge, MA, Belkap Press of Harvard Univ. Press.

COOLICAN, H. (1994) *Research Methods and Statistics in Psychology.* London, Hodder and Stoughton.

COOPER, K. and OLSON, M. (1996) The multiple 'I's' of teacher identity: in: M.Kompf, T.Boak, W.R.Bond and D.Dworet (Eds.) *Changing research and practice: teachers' professionalism, identities and knowledge.* London, Falmer Press.

CRESSWELL, M. J (1990) Gender effects in GCSE – some initial analyses. Paper prepared for *Nuffield Seminar* at University of London, Institute of Education, 29 June 1990.

CRESSWELL, M. J (1996) Defining, Setting and Maintaining Standards in Curriculum Embedded Examinations: Judgemental and Statistical Approaches in H. Goldstein and T. Lewis (1996) *Assessment: Problems, Developments and Statistical issues* London, Wiley.

CRESSWELL, M. J (1997) *Examining Judgements: Theory and Practice of Awarding Public Examination Grades.* Unpublished PhD thesis, University of London, Institute of Education.

CRESSWELL, M. J. and GIBB, J. (1987) *The Second International Mathematics Study in England and Wales.* Windsor, NFER – Nelson.

CRESSWELL, M. J. and HOUSTON, J.G. (1991) Assessment of the National Curriculum – some fundamental considerations. *Educational Review*, 43, 63 – 78.

DAILY EXPRESS, untitled article on iGCSE and GCSE, 3 September 2005, p. 9.

DAILY TELEGRAPH, 2004 GCSE results, 27 August 2004, pp. 6-7

DAY, C., KINGTON, A., STOBART, G. and SAMMONS, P. (2006) The personal and professional selves of teachers: stable and unstable identities, *British Educational Research Journal*, Vol. 32, No. 4, pp. 601-616.

DEPARTMENT OF EDUCTION AND SCIENCE (DES) (1974) *Educational Disadvantage and the Educational Needs of Immigrants (Cmnd5720).* London, HMSO.

DEPARTMENT OF EDUCTION AND SCIENCE (DES) (1980) Letter (28 February, 1980) from the DES to the examining boards inviting them to start work on preparing national criteria for twenty subjects. London, DES.

DEPARTMENT OF EDUCTION AND SCIENCE (DES) (1982) *Examinations at 16-plus: a statement of policy.* London, HMSO.

DEPARTMENT OF EDUCTION AND SCIENCE (DES) / WELSH OFFICE (WO) (1985) *General Certificate of Secondary Education: the National Criteria.* London and Cardiff, HMSO.

DEPARTMENT OF EDUCTION AND SCIENCE (DES) / WELSH OFFICE (WO): HER MAJESTY'S INSPECTORATE (HMI) (1988) *The General Certificate of Secondary Education: an interim report on the introduction of the new examination in England and Wales.* London and Cardiff, HMSO.

DEPARTMENT OF EDUCTION AND SCIENCE (DES) /WELSH OFFICE (WO) (1989) *Science in the National Curriculum.* London and Cardiff, HMSO.

DEPARTMENT OF EDUCTION AND SCIENCE (DES) /WELSH OFFICE (WO) (1993) *Science Teacher Retention.* London, Cardiff: HMSO.

DORE, R. (1976) *The Diploma Disease.* London, Unwin.

DREW, D. and GRAY, J (1990) The fifth year examination achievements of black young people in England and Wales. *Educational Research*, 32(3), pp. 107-17.

DREW, D. and GRAY, J (1991) The black-white gap in examination results: a statistical critique of a decades's research. *New Community*, 17(2), pp. 159-72.

ECKSTEIN, M. A. and NOAH, H. J. (1992) (Eds) *Examinations: Comparative and International Studies*. Oxford, Pergamon Press.

EGGLESTON, J. (1990) School Examinations – Some Sociological Issues. In Horton, T. (Ed.) *Assessment Debates*. Milton Keynes: The Open University.

ELWOOD, J. (1995) Undermining Gender Stereotypes: examination and coursework performance in the UK at 16. *Assessment in Education*, Vol.2, No.3, 1995.

ELWOOD, J. (2001) Examination Techniques: Issues of Validity and Effects on Pupils' Performance. In Scott, D. (Ed.) *Curriculum and Assessment*. Westport, Ablex Publishing.

ELWOOD, J. and COMBER, C. (1996) *Gender Differences in Examinations at 18+: Final Report.* London, Institute of Education, University of London.

ELWOOD, J. and MURPHY, P. (2002) Tests, tiers and achievement: gender and performance at 16 and 14 in England. *European Journal of Education*. 37, 4, pp. 395-416.

ELY, M., ANZUL, M., FRIEDMAN, T., GARNER, D. and McCORMACK STEINMETZ, A. (1991) *Doing Qualitative Research: Circles Within Circles.* London, Routledge.

ERCIKAN, K. and ROTH, W. (2006) What good is polarizing research into qualitative and quantitative? *Educational Researcher*, Vol.35, No. 5, pp. 14-23.

EYSENCK, H.J.(1973) *The Measurement of Intelligence.* Baltimore, Williams and Wilkins.

FILER, A. (2000) (Ed.) *Assessment: Social Practice and Social Product.* London, RoutledgeFalmer.

FIRESTONE, W.A. (1989) Educational policy as an ecology of games. *Educational researcher*, 18 (7).

FIRESTONE, W.A. (1998) A tale of two tests: tensions in assessment policy. *Assessment in Education*, Vol.5, No.2, 1998.

FITZ-GIBON, C. T. and VINCENT, L. (1994) *Candidates' performance in public examinations in Mathematics and Science* London, School Curriculum and Assessment Authority.

FOGELIN, R. J., (1967) *Evidence and Meaning: studies in analytic philosophy.* London, Routledge.

FOREST, G.M. and SHOESMITH, D. (1985) *A second review of comparability studies* Manchester, Joint Matriculation Board.

FOUCAULT, M. (1977) *Discipline and Punish: the Birth of Prison.* New York, Vintage Books.

FRENCH, S., SLATER, J.B., VASSILOGLOU, M. and WILLMOTT, A.S. (1987) *Descriptive and Normative Techniques in Examination Assessment.* Oxford, UODLE

GILL, J. (1994) *Differences in the making: the construction of gender in Australian schooling.* Unpublished PhD thesis. Adelaide, University of Adelaide.

GILLBORN, D. and GIPPS C. (1996) *Recent research on the achievements of ethnic minority pupils.* London, OFSTED.

GILLBORN, D. and YOUDELL, D. (1998) *Ethnic Origin and Selection in GCSE English and Mathematics: Final Report.* London, University of London, Institute of Education.

GIPPS, C. (1989) The Debate Over Standards and the Uses of Testing. In B. Moon, P. Murphy and S. Raynor (Eds.) *Policies for the Curriculum.* London, Hodder and Stoughton.

GIPPS, C. (1990) *Assessment: A Teachers Guide to the Issues.* London, Hodder and Stoughton.

GIPPS, C. (1994) *Beyond Testing.* Lewes, Falmer Press.

GIPPS, C., MACINTOSH, H., TORRANCE, H., MURPHY, R., GOLDSTEIN, H. and NUTTALL, D. (1986) *The GCSE: An Uncommon Examination.* London, University of London, Institute of Education.

GIPPS, C. and MURPHY, P. (1994) *A Fair Test.* Buckingham, Open University Press.

GLASER, R. (1963) Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 18, pp. 519-21.

GLICK, P., WILK, K. and PERREALT, M. (1995) Images of occupations: components of gender and status in occupational stereotypes. *Sex Roles,* 32. pp. 565-582.

GOACHER, B. (1984) *School Reports to Parents.* Florence, KY, Thomson Delmar Learning.

GOLDSTEIN, H. (1986) Models for equating test scores and for studying the comparability of public examinations. In Nuttall, D. L. (Ed.) *Assessing Educational Achievement.* London, Falmer Press.

GOLDSTEIN, H. (1991) Assessment in Schools: an alternative framework. London, Institute for Public Policy Research.

GOLDSTEIN, H. (1995) *Multilevel Statistical Models.* London, Edward Arnold.

GOLDSTEIN, H., RASHBASH, J., YANG, M., WOODHOUSE, G., PAN, H., NUTTALL, D. and THOMAS, S. (1993) A Multilevel Analysis of School Examination Results. *Oxford Review of Education,* 19(4), pp. 425-433.

GOMM, R., HAMMERSLEY, M. and FOSTER, P. (Eds.) (2000) *Case Study Methods.* Thousand Oakes, CA, Sage.

GOOD, F. J. (1986) *Differentiated Assessment: some problems for examiners.* London, Secondary Examinations Council.

GOOD, F. J. (1989) Setting Common Examination Papers that Differentiate. *Educational Studies,* Vol.15, No.1.

GOOD, F. J. and CRESSWELL, M. J. (1988a) *Differentiated Assessment: grading and related issues.* London, Secondary Examinations Council.

GOOD, F. J. and CRESSWELL, M. J. (1988b) Grade awarding judgements in differentiated examinations. *British Educational Research Journal,* 14, 261-279.

GOOD, F. J. and CRESSWELL, M. J. (1988c) Placing candidates who take differentiated papers on a common grade scale. *Educational Research*, Vol. 30, No. 3, 1988.

GOOD, F. J. and CRESSWELL, M. J. (1988d) Can Teachers Enter Candidates appropriately for Examinations Involving Differentiated Papers? *Educational Studies*, Vol. 14, No. 3, 1988

GOODSON, I.F. and HARGREAVES, A. (Eds.) (1996) *Teachers' Professional Lives*. London, Falmer Press.

GORARD, S., SALISBURY J. and REES G. (1999) *Revisiting the apparent underachievement of boys: reflections on the implications for educational research.* British Educational Research Association Annual Conference, University of Sussex in Brighton, September 2 -5 1999.

GOVERNMENT CIRCULAR (1965) 10/65. London, HMSO.

GREENO J.G., PEARSON P.D. and SCHOENFELD A.H. (1998). Achievement and Theories of Knowing and Learning. In R. McCormick, and C. Paechter, (Eds.) *Learning and Knowledge*, pp.136-153. London, Paul Chapman Educational Publishing.

GUBA, E.G. and LINCOLN, Y.S. (1981) *Effective evaluation: improving the usfulness of evaluation results through responsive and naturalistic approaches.* San Fransisco, CA, Jossey-Bass.

GUBA, E.G. and LINCOLN, Y.S. (1982) Epistemological and methodological bases of naturalistic inquiry. *Educational Communication and Technology Journal*, 30(4) pp. 233-252.

GUBA, E.G. and LINCOLN, Y.S. (1994). Competing Paradigms in Qualitative Research. In Denzin N.K. and Lincoln Y.S. (Eds.) *Handbook of Qualitative Research.* Thousand Oaks, CA, Sage.

GUBA, E.G. and LINCOLN, Y.S. (1998). Competing Paradigms in Qualitative Research: the Landscape of Qualitative Research. CA, Sage.

GUSKEY,T.R. and KIFER, E.W. (1989) *Ranking School Districts on the Basis of Statewide Test Results: Is It Meaningful or Misleading?* San Francisco: American Educational Research Association.

HADOW REPORT (1926) *The Education of the Adolescent.* Report of the Board of Education Consultative Committee.

HALSEY, A.H. and GARDNER, L. (1953) Selection for secondary education and achievement in four grammar schools. *British Journal of Sociology*, 4, pp. 60-75.

HAMMERSLEY, M. (1989) *The Dilemma of Qualitative Method.* London, Routledge.

HAMMERSLEY, M. and ATKINSON, P. (1995) *Ethnography: principles in practice, second edition.*, London, Routledge.

HANSON, F.A. (2000) How tests create what they are intended to measure. In A. Filer *Assessment: Social Practice and Social Product.* London, Routledge Falmer.

HARGREAVES, A. (1994) *Changing teachers, changing times.* London, Falmer Press.

HEBB, D. O. (1949) *The Organisation of Behaviour.* New York, John Wiley.

HILDEBRAND, D.K., LANGE, J.D. and ROSENTHAL, H. *Prediction Analysis of Cross-tabulations.* New York, Wiley.

HIRSCH, N. D. M. (1928) An experimental study of East Kentucky Mountaineers: A study in heredity and environment. *Genet Psychol Monogr*, 1928, 3, 183-244.

HORTON, T. (1990) (Ed.) *Assessment Debate.* London, Hodder and Stoughton.

HOUSTON, J.G. (1980) *Report of the Inter-board Cross-moderation Study in English Literature at Ordinary level: 1975.* Aldershot, Associated examining Board.

HYCNER, R.H. (1985) Some guidelines for the Phenomenological Analysis of Interview Data, *Human Studies.* 8 (1985) pp. 279-303.

INDEPENDENT (1998) Classroom rescue for Britain's lost boys. *The Independent.* 5/1/98, p. 8.

INGENKAMP, K. (1977) *Educational Assessment.* Slough, NFER.

INTER-GROUP RESEARCH COMMITTEE (IGRC) (1993) *Inter-group statistical reports of GCSE (UK) science examinations.* Cambridge, University of Cambridge Local Examinations Syndicate for all GCSE Groups.

INTERNATIONAL BACCALAUREATE ORGANISATION (IBO) (1994) (2004) *Examination Entries.* Cardiff, International Baccalaureate Organisation.

IRESON, J. and HALLAM, S. (2001) *Ability grouping in education.* London, Sage.

JAMES-WILSON, S. (2001) *The influence of ethnocultural identity on emotions and teaching.* Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, April 2000.

JOHNSON, S. and COHEN, L. (1983) *Investigating Grade Comparability through Cross-moderation.* London, Schools Council.

JOHNSON, S. and MURPHY, P. (1986) *Girls and Physics: Reflections on APU Survey Findings.* APU Occasional Paper No.4. London, DES.

JOHNSON, R.B. and ONWUEGBUZIE, A.J. (2004) Mixed methods research: a research paradigm whose time has come. *Educational Researcher*, 33:7, pp. 14-26.

JOINT COUNCIL for NATIONAL CRITERIA (1981) *Glossary of Terms for a single System of Examining at 16-plus.*

JOSEPH, K. (1984a) Speech at the North of England Education Conference, Sheffield, 6 January, DES Press Release 1/84.

JOSEPH, K. (1984b) Speech to Assistant Masters and Mistresses Association, Bournemouth, April, DES Press Release 65/84.

KELCHTERMANS, G. and VANDENBERGHE, R. (1994) Teachers' professional development. A biographical perspective. *Journal of Curriculum Studies*, 26(1), pp. 45-62.

KELLY, A. (1976) A study of the comparability of external examinations in different subjects. *Research in Education*, 16, pp. 37-63.

KERLINGER, F.N (1970) *Foundations of behavioural research.* NY Holt, Rinehart and Winston.

KINGDON, M. and STOBART, G. (1988) *GCSE Examined.* Lewes, Falmer Press.

LANDIS, J.R. and KOCH, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, 33, pp. 159-174.

LAPHAM, C. (1989, 1992) *Chief Examiner's Report: Science.* Cardiff, Welsh Joint Education Committee.

LAWSON, J. and SILVER, H. (1973) *A Social History of Education in England.* London, Methuen.

LAWTON, D. (1980) *Politics of the School Curriculum.* London, Routledge and Kegan Paul.

LAVE, J. (1988) *Cognition in Practice.* CA, CambridgeUniv. Press.

LAVE, J. and WENGER, E, (1991) *Situated Learning: Legitimate Peripheral Participation.* Cambridge, University Press.

LITTLE, A. and WOLF, A. (1996) (Eds) *Assessment in Transition: learning, monitoring, and selection in international perspective*, Oxford, Elsevier Science Ltd.

MATTHEWS, J. C. (1985) *Examinations.* London, Allen and Unwin.

MAYKUT, P. and MOREHOUSE, R. (1994) *Beginning Qualitative Research: A Philosophic and Practical Guide.* Lewes, Falmer Press.

MERRIAM, S.B. (1998) *Qualitative research and case study application in education* (Rev. ed.). San Fransisco, CA, Jossey-Bass.

MESSICK, S. (1988) 'Validity', in Linn, R. (Ed.) *Educational Measurement* (3rd. edn. Pp. 13-103). American Council on Education, Washington, Macmillan.

MESSICK, S. (1989). Meaning and values in test validation: the science and ethics of assessment, *Educational Researcher*, 18.

MIDLAND EXAMINING GROUP (MEG) (1995) *A Comparability Study in GCSE English: a study based on the summer 1994 examinations.* Cambridge, organized by Midland Examining Group on behalf of the Inter-Group Research Committee.

MINICK, N. (2005) *The development of Vygotsky's thought: an introduction to thinking and speech.* Newbury Park, CA, Sage.

MONTGOMERY, R. (1965) *Examinations. An Account of their Evolution as Administrative Devices in England.* London, Longmans.

MONTGOMERY, R. (1978) *A New Examination of Examinations.* London, Routledge and Kegan Paul.

MORTIMORE, P. and WHITTY, G. (1997) *Can School Improvement Overcome the Effects of Disadvantage?* London, Institute of Education.

MURPHY, P (2000). Equity Assessment and Gender in J. Salisbury and S. Ridell (Eds.), *Gender, Policy and Educational Change.* London, Routledge.

MURPHY, P. and ELWOOD, J. (1998) Achievement: exploring the link. *International journal of inclusive education*, 2:22, pp. 95-118, Taylor & Francis, 1998.

MURPHY, P & IVINSON G. (2004). Gender Differences in Educational Achievement: A Socio-cultural Analysis. In M. Olssen (Ed.) *Culture and learning: access and opportunity in the classroom.* USA, Information Age Publishing.

MURPHY, P. and WHITELEGG, E. (2006). *Girls in the Physics Classroom: a review of the research on the participation of girls in physics.* Milton Keynes, The Open University.

MURPHY, R. (1979) Teachers' assessments and GCE results compared, *Educational Research*, 22, 54-59.

MURPHY, R. (1981))-level grades and teachers' estimates as predictors of the A-level results of UCCA applicants, *British Journal of Educational Psychology*, 51, 1-9.

MURPHY, R. (1982) Sex differences in objective test performance. *British Journal of Educational Psychology*, 52, pp. 213-219.

MURPHY, R. (1986) Grade criteria and the GCSE, in GIPPS, C., MACINTOSH, H., TORRANCE, H., MURPHY, R., GOLDSTEIN, H. AND NUTTALL, D. (1986) *The GCSE: An Uncommon Examination.* London, University of London Institute of Education.

MURPHY, R., BURKE, P., COTTON, T., HANCOCK, J., PARTINGTON, J., ROBINSON, C., TOLLEY, H., WILMUT, J. and GOWER, R. (1996) *The Dynamics of GCSE Awarding: report of a project conducted for the School Curriculum and Assessment Authority.* London, SCAA.

MURPHY, R. and TORRANCE, H. (1990) The Need for Change, in HORTON, T. (Ed) *Assessment Debates.* Milton Keynes, The Open University.

NATIONAL ASSOCIATION OF SCHOOLMASTERS / UNION OF WOMEN TEACHERS (NAS/UWT). Union members briefing paper on the introduction of GCSE. NAS/UWT.

NESPOR, J. (1997) *Tangled up in school. Politics, space, bodies and signs in the educational process.* NJ, Lawrence Erlbaum Ass.

NEWBOLD, C. A. (1994) *Personal communication.*

NEWBOLD, C. A. (1995) *Personal communication.*

NEWBOLD, C. A and MASSEY, A.J. (1979) *Comparability using a common element.* Cambridge, University of Cambridge Local Examinations Syndicate (UCLES).

NEWBOLD, C.A. and SCANLON, L.A.J. (1981) *An Analysis of Interaction between Sex of Candidate and Other Factors.* Cambridge, TDRU.

NEWTON, P., BAIRD, J-A, GOLDSTEIN, H.PATRICK, H.and TYMMS, P. (2008) *Techniques for monitoring the comparability of examination standards.* London, Qualifications and Curriculum Authority (QCA).

NIAS, J. (1989) *Primary teachers talking.* London, Routledge and Kegab Paul.

NIAS, J. (1996) Thinking about felling: the emotion sin teaching, *Cambridge Journal of Education,* 26(3), pp.293-306.

NOAH, H. J. and ECKSTEIN, M. A. (1989) Tradeoffs in Examination policies: an International Comparative Perspective. *Oxford Review of Education*, 15(1), 17-27.

NOAH, H. J. and ECKSTEIN, M. A. (1992) *Examinations: Comparative and International Studies* (Comparative and International Education Series). Oxford, Pergamon Press.

NORTHERN EXAMINATIONS and ASSESSMENT BOARD (NEAB) (1993) *Personal communication.*

NORTHERN EXAMINATIONS and ASSESSMENT BOARD (NEAB) (1994) *Personal communication.*

NUTTALL, D. L. (1971) *The 1968 CSE Monitoring Experiment*, (Schools Council Working Paper 34). London, Evans/Methuen Educational.

NUTTALL, D. L. (1990) The GCSE: promise vs. reality, in P. Broadfoot, R. Murphy. and H. Torrance (1990) *Changing Educational assessment: International Perspectives and Trends*. London, Routledge.

NUTTALL, D. (1986) (Ed.) *Assessing Educational Achievement*. London, Falmer Press.

NUTTALL, D.L., BACKHOUSE, J.K. and WILMOTT, A.S. (1974) *Comparability of Standards Between Subjects*, Schools Council Examinations Bulletin 29. London, Evans/Methuen Educational.

NUTTALL, D. L., GOLDSTEIN, H., PROSSER, R., and RASBASH, J. (1989) Differential School Effectiveness, *International Journal of Educational Research*, 13, 769-76.

ONWUEGBUZIE, A.J. (2002) Positivists, post-positivists, post-structuralists and post-modernists: Why can't we all get along? Towards a framework for unifying research paradigms. *Education*, 122(3),, pp. 518-530.

ORR, L. and NUTTALL, D.L. (1983) *Determining standards in the proposed single system of examining at 16+*. London, Schools Council.

OXFORD CERTIFICATE OF EDUCATIONAL ACHIEVEMENT (OCEA) (1985) *The Oxford Certificate of Educational Achievement Teachers' Guide*. Oxford, Oxford International Assessment Services.

OXFORD UNIVERSITY DELEGACY OF LOCAL EXAMINATIONS (OUDLE) (1995) *Personal communication.*

PARLIAMENTARY PAPERS (1898) *Report of the Departmental Committee on Defective and Epileptic Children*, 1898 xxvi, 752.

PATERSON, L. and GOLDSTEIN, H. (1991) New statistical methods for analysing social structures: an introduction to multilevel models. *British Educational Journal*, 17, 387-94.

PATTON, M.Q. (1990) *Qualitative evaluation and research methods* (2nd. Edition). Newbury Park, CA, Sage.

PETCH, J. A. (1964) *School Estimates and Examination Results Compared.* Manchester, Joint Matriculation Board.

PIAGET, J. (1950) *The Psychology of Intelligence*. London, Routledge.

PLOWDEN REPORT (1967) *Children and their Primary Schools: A Report of the Central Advisory Council for Education.* London, HMSO.

POLLITT, A., ENTWISTLE, N., HUTCHINSON, C. and DELUCA, C. (1985) *What Makes Exam Questions Difficult?* Edinburgh, Scottish Academic Press.

POPHAM, W. J. (1987) The merits of measurement-driven instruction, *Phi Delta Kappan*, 68, 680 – 82.

POPHAM, W. J. and SIROTNIK, K.A. (1973) *Educational statistics: use and interpretation.* London, Harper Row.

PREECE, P.F.W., SKINNER, N.C. and RIALL, R.A.H. (1999) The gender gap and discriminating power in the National Curriculum Key Stage three science assessments in England and Wales. *International Journal of Science Education.* 21 (9) pp. 978-987.

PRING, R. (1984) Confidentiality and the right to know. In C.Aldeman (Ed.) *The politics and ethics of evaluation.* London, Groom Helm.

PRYOR, J. and TORRANCE, H. (2000) Questioning the Three Bears. *Assessment: Social Practice and Social Product.*

QUINLAN, M. (1993) *Delta Index* Paper given at the inter-group research Committee seminar on Interpreting Examination Statistics held at the offices of the University of London Examinations and assessment Council, March, 1993.

RADNOR, H. (1987) *GCSE: the Impact of the Introduction of GCSE at LEA and School Level,* National Foundation for Educational Research.

RATCLIFFE, P. (1994) *A Comparability Study in GCSE Geography: a study based on the Summer 1993 examinations.* Manchester, organized by Northern Examinations and Assessment Board on behalf of the Inter-Group Research Committee.

RESNICK, L. B. (1976) (Ed) *The Nature of Intelligence.* New York, John Wiley.

REYNOLDS, C. (1996) Cultural scripts for teachers: identities and their relation to workplace landcapes, in M.Kompf, T.Boak, W.R.Bond and D.Dworet (Eds.) *Changing research and practice: teachers' professionalism, identities and knowledge.* London, Falmer Press.

ROBINS, C. (1972) *Comparability Studies.* Cardiff, Welsh Joint Education Committee.

ROBINSON, P. and OPPENHEIM, C. (1998) *Social Exclusion Indicators.* London, Institute of Public Policy Research.

ROGOFF, B. (1995) Observing sociocultural activity on three planes: participatory appropriation, guided participation, and apprenticeship. In J.V. Wertsch, P.del Rio and A.Alvarez (Eds.), *Sociocultural studies of mind* (pp.139-164). New York, Cambridge Univ. Press.

RUDDOCK, G. J., TOMLINS, B., MASON, K., HOLDING, B., REISS, M., KEYS, M., FOXMAN, D. and SCHAGEN, I. (1993) *Evaluation of National Curriculum Assessments at Key Stage 3: report on the 1992 national pilot assessment in mathematics and science,* unpublished research report. London, SEAC.

RUDDOCK, G. J., STURMAN, L., SCHAGEN, I., STYLES, B., GNALDI, M and VAPPULA, H. (2003) *Where England stands in Trends in International Mathematics and Science Study (TIMMS) 2003: Summary of national report for England.* NFER Department for Education and Skills 12.

RIDING, R. and BUTTERFIELD, S. (1990) *Assessment and Examination in the Secondary School.* London, Routledge.

SADLER, D. R. (1985) The origins and functions of evaluative criteria, *Educational Theory*, 35, pp. 285 – 297.

SADLER, D. R. (1987) Specifying and promulgating achievement standards. *Oxford Review of Education*, 13, 191 – 209.

SADLER, D. R. (1989) Formative assessment and the design of instructional systems, *Instructional Science*, 18, 119 – 144.

SAMMONS, P., NUTTALL, D. CUTTANCE, P. and THOMAS, S. (1995) Continuity of school effects: a longitudinal analysis of primary and secondary school effects on GCSE performance. *School Effectiveness and Improvement.* 6. pp. 285-307.

SCHOOL CURRICULUM AND ASSESSMENT AUTHORITY (SCAA) (1993a) *GCSE Regulations and Criteria*, London, The School Curriculum and Assessment Agency.

SCHOOL CURRICULUM AND ASSESSMENT AUTHORITY (SCAA) (1993b) *GCSE Mandatory Code of Practice*, London, The School Curriculum and Assessment Agency.

SCHOOL CURRICULUM AND ASSESSMENT AUTHORITY (SCAA) (1995a) *GCSE Regulations and Criteria*, London, The School Curriculum and Assessment Agency.

SCHOOL CURRICULUM AND ASSESSMENT AUTHORITY (SCAA) (1995b) *GCSE Mandatory Code of Practice*, London, The School Curriculum and Assessment Agency.

SCHOOL CURRICULUM AND ASSESSMENT AUTHORITY (SCAA) (1996) *Tiering in GCSE Examinations: a guide for teachers*, London, Schools Curriculum and Assessment Authority.

SCHOOLS COUNCIL (1966) *The 1965 CSE Monitoring Experiment* (Working Paper 6), parts I and II. London, HMSO.

SCHOOLS COUNCIL (1971) *A Common System of Examining at 16+.* Examinations Bulletin 23. London, HMSO.

SCHOOLS COUNCIL (1975) *The Whole Curriculum 13-16.* Working Paper 53. London, HMSO

SCHOOLS COUNCIL (1979) *Standards in Public Examinations; Problems and Possibilities*, Report from the Schools Council Forum on Comparability. London, HMSO.

SCOTTISH EDUCATION DEPARTMENT (SED) (1986) *Assessment in Standard Grade Courses: proposals for simplification* (McClelland Report). Edinburgh, SED.

SECONDARY EXAMINATIONS COUNCIL (SEC) (1984) *The Development of Grade-Related Criteria for the General Certificate of Secondary Education. A briefing paper for working parties.* London, Secondary Examinations Council.

SECONDARY EXAMINATIONS COUNCIL (SEC) (1985) *Working Paper Two – Coursework Assessment in GCSE.* London, Secondary Examinations Council.

SFARD, A. (1998) On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, 27(2), pp. 4-13.

SIKES, P.J., MEASOR,L. and WOODS, P. (1991) Berufslaufbhan und Identitat im Lehrerberuf, in: E. Terhart (Ed.) Unterrichten als beruf, pp. 231-248 (Cologne, Bohlau) cited in D. Beijaard (1995) Teachers' prior experiences and perceptions of professional identity, *Teachers and Teaching*, 1(2), pp. 281-294.

SIMON, B. (1953) *Intelligence Testing and the Comprehensive School.* London, Lawrence and Wishart.

SKURNIK, L.S. and HALL, J. (1969) *The 1966 CSE Monitoring Experiment*, (Schools Council Working Paper 21). London, HMSO.

SKURNIK, L.S. and CONNAUGHTON, I. M.. (1970) *The 1967 CSE Monitoring Experiment*, (Schools Council Working Paper 30). London, Evans/Methuen Educational.

SLEEGERS, P. and KELCHERMANS, G. (1999) Inleiding op het themanummer: professionele identeit van leraren [Professional identity of teachers], *Pedagogisch Tijdschrift*, 24, pp. 369-374.

SMITH, D. and TOMLINSON, S. (1989) *The School Effect.* London, Policy Studies Institute.

SOUTHERN EXAMINING GROUP (SEG) (1995) Private verbal communication with Ann Benson, Computing and Statistics Officer. Guildford, Southern Examining Group.

SOUTHERN EXAMINING GROUP (SEG) (1995) *A Comparability Study in GCSE Science: a study based on the summer 1994 examinations.* Guildford, organized by Southern Examining Group on behalf of the Inter-Group Research Committee.

SOUTHERN EXAMINING GROUP (SEG) (1996) Private verbal communication with Ann Benson, Computing and Statistics Officer. Guildford, Southern Examining Group

SOWELL, D. (1970) CSE-Grades and Teachers' Forecasts. *Educational Research*, 13, 28-35.

SPEARMAN, C. (1927) *The Nature of 'Intelligence' and the Principles of Cognition.* London, Macmillan.

SRAC (1976) *Comparative Statistics Charts, 1975.* Unpublished, Standing Research Advisory Committee of the GCE Boards.

STAKE, R.E. (1978) The case study method in social enquiry. *Educational Researcher* Feb. 1978, 7(2), pp. 5-8

STAKE, R.E. (1994) Case Studies. In N.K.Denzin and Y.S Lincoln *Handbook of Qualitative Research.* Thousand Oaks, CA, Sage.

STENHOUSE,, L. (1978) Case Study and Case Records: towards s contemporary history of education. *British Educational Research Journal*, 4 (2) pp. 21-39.

STERNBERG, R. J. (1998) *In Search of the Human Mind.* Fort Worth, Harcourt Brace College Publisher.

STOBART, G. (1989) *A Comparability Study in GCSE History: a study based on the work of candidates in the Summer 1988 examinations*. London, organized by University LEAG on behalf of the Inter-group Research Committee for the GCSE.

STOBART, G., WHITE, J., ELWOOD, J., HAYDEN, M. and MASON, K. (1992). *Differential Performance in Examinations at 16 plus: English and Mathematics*. London, Schools Examination and Assessment Council.

STOBART, G., ELWOOD, J., JANI, A. and QUINLAN, M. (1994) *A Comparability Study in GCSE History: a study based on the Summer 1993 examinations*. London, organized by University of London Examinations and Assessment Council on behalf of the Inter-group Research Committee for the GCSE.

SUMISON, J. (2002) Becoming, being and unbecoming an early childhood educator: a phenomenological case study of teacher attrition. *Teaching and Teacher Education*. 18, pp. 869-885.

SUTHERLAND, G. (1984). *Ability, Merit and Measurement: Mental Testing and English Education 1880-1940*. Oxford, Clarendon Press.

SUTHERLAND, G. (1996). Assessment: Some Historical Perspectives. In H. Goldstein and T. Lewis (Eds.), *Assessment: Problems, Developments and Statistical Issues*. Chichester, John Wiley & Sons.

TATTERSALL, K. (1983) Differentiated examinations: a strategy for assessment at 16+? *Schools Council Bulletin 42*. London, Methuen.

TATTERSALL, K. (1994) The Role and Functions of Public Examination. *Assessment in Education*, Vol. 1, No. 3, 1994.

TECHNIQUEST (1995) Personal communication with the Director of Techniquest, Cardiff.

THORNDIKE, E. L. (1933) The effect of interval between test and retest on the constancy of the IQ. *Journal of Educational Psychology*, 1933, 24, 543-549.

THORNDIKE, E. L., BREGMAN, E. O. and COBB, M. V. (1927) *The Measurement of Intelligence*. New York, Teachers College, Columbia University

THURSTONE, L.L. (1938) *Primary Mental Abilities*. Chicago, University of Chicago Press.

TIMES EDUCATIONAL SUPPLEMENT (TES) (1991) 'GCSE fails to eliminate inequality', *Times Educational Supplement* 15 February, p.1.

TIMES EDUCATIONAL SUPPLEMENT (TES) (1993) Article on School League Tables, *Times Educational Supplement* 10 September, p 12.

TIMES EDUCATIONAL SUPPLEMENT (TES) (1994) Article by Smithers, A., *Times Educational Supplement* 25 March, p. 10.

TIMES EDUCATIONAL SUPPLEMENT (TES) (1995) 'Standards Tumble', *Times Educational Supplement* 25 August 1995, p. 10.

TIMES EDUCATIONAL SUPPLEMENT (TES) (2005) Boys are getting better ... but so are girls, *Times Educational Supplement* 17 June 2005, p. 8.

TIMES EDUCATIONAL SUPPLEMENT (TES) (2005) Good Schools Guide Supplement, *Times Educational Supplement* 26 August 2005.

TIMSS (1999) and (2003) Trends in International Mathematics and Science Studies. TIMSS.

TORRANCE, H. (2007) Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education* Vol.14, No. 3, November 2007, pp. 281-291

TROYNA, B. (1991) Underachievers or under-rated? The experience of pupils of South African origin in a secondary school. *British Educational Research Journal*, 17(4), pp. 359-74

TUCKMAN, B.W. (1972) *Conducting Educational Research.* New York, Harcourt Brace Jovanovich.

UNGER, J. (1984) The Historical Background: Examinations and controls in Pre-Modern China. *Comparative Education,* 20, pp. 7-26.

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE (UCLES) (1993) *Personal communication*

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE (UCLES) (1994) *Personal communication*

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE (UCLES) (2001) *Personal communication*

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE GROUP (2005) *Personal communication.*

UNIVERSITY OF LONDON EXAMINATIONS AND ASSESSMENT COUNCIL (ULEAC) (1995) *A Comparability Study in GCSE Mathematics: a study based on the summer 1994 examinations,* London, organized by University of London Examinations and Assessment Council on behalf of the Inter-Group Research Committee.

UNIVERSITY OF OXFORD (1852) Report of the Oxford University Commission, *Examinations,* 1852, Oxford.

VERNON, P.E. (1950) *The Structure of Human Abilities*, London, Methuen New York: Wiley.

VERNON, P.E. (1957) *Secondary School Selection*, London, Methuen.

VERNON, P.E. (1979) *Intelligence Testing 1928-1978 What Next?* Edinburgh, The Scottish Council for Research in Education.

VIDICH, J.A. and BENSMAN, J. (1968) cited in Peshkin, (1993) The Goodness of Qualitative Research, *Educational Researcher* 22(2), pp. 23-29.

VOIGT, J. (1985) Patterns and routines in classroom interaction. *Recherches en Didactique des Mathematique.* 6, pp. 69-118.

VON GLASERSFELD (1989) Constructivism. In T. Husen and T.N. Postlethwaite (Eds.) *The international encyclopedia of education* 1st. ed., supplement vol. 1, pp. 162-163. Oxford, Pergamon.

VYGOTSKY, L.S. (1960) *Razvitie vysshikh psikhicheskikh funkstii [The development of the higher metal functions].* Moscow, Akad. Ped. Nauk. RSFSR.

VYGOTSKY, L.S. (1962) *Thought and Language.* London, Wiley.

VYGOTSKY, L.S. (1978) *Mind in Society: the Development of Higher Psychological Processes.* (Ed). M.Cole, V. John-Steiner, S.Schreibner, E Souberman. Cambridge, MA, Harvard Univ. Press.

VYGOTSKY, L.S. (1981) The instrumental method in psychology. In *The Concept of Activity in Soviet Psychology.* (Ed.) J.Wertsch, pp. 3-35. Armonk,NY, Sharpe.

WADDELL REPORT (1978) *Secondary Schools Examinations; a single system at 16-plus.* London, HMSO.

WALKERDINE, V. (1989) Developmental Psychology and Pedagogy in P. Murphy and B. Moon (Eds.) *Developments in Learning and Assessmen.,* London, Hodder and Stoughton.

WARWICK, D.P. and OSHERSON, S. (1973) (Eds.) *New Directions in Quantitative Comparative Sociology.* Englewood-Cliffs, N.J.,Prentice Hall.

WELSH JOINT EDUCATION COMMITTEE (WJEC) (1993) *Personal communication.*

WELSH JOINT EDUCATION COMMITTEE (WJEC) (1994) *Personal communication.*

WELSH JOINT EDUCATION COMMITTEE (WJEC) (1995) *Personal communication.*

WELSH JOINT EDUCATION COMMITTEE (WJEC) (1995) *A Comparability Study in GCSE Science: a study based on the summer 1994 examinations,* Cardiff, organized by Welsh Joint Education Committee on behalf of the Inter-Group Research Committee

WENGER, E. (1998) *Communities of Practice, Learning, Meaning, and Identity.* Cambridge, Cambridge Univ. Press.

WERTSCH, J.V. (1991). *Voices of the mind: A socio-cultural approach to mediated action.* Cambridge, MA, Harvard University Press.

WERTSCH, J. V., del RIO, P. and ALVAREZ A. (1995) *Sociocultural Studies of Mind (Learning in Doing: Social, Cognitive and Computational Perspectives).* San Fransisco, CA, Wiley and Sons.

WERTSCH, J.V.and TULVISTE, P. (2005) L.S.Vygotsky and contemporary developmental psychology. In H.Daniels (Ed.) *An Introduction to Vygotsky.* Hove, Routledge.

WILIAM, D. (1996) Standards in examinations: a matter of trust. *Curriculum Journal,* Vol. 7, Iss. 3, pp293-306.

WILIAM, D. and BARTHOLOMEW, H. (2004) It's not which school but which set you're in that matters: the influence of ability grouping practices on student progress in mathematics. *British Educational Research Journal,* Vol. 30, No. 2, pp. 279-291.

WILLINGHAM, W.W. and COLE, N.S (1997) *Gender and Fair Assessment.* Mahwah, NJ, Lawrence Erlbaum Associates.

WILLMOTT, A.S. (1977) *CSE and GCE Grading Standards: the 1973 Comparability Study,* London, Macmillan Education Ltd.

WILLMOTT, A.S. (1980) Twelve Years of Examination Research ETRU 1963-1977. London, Schools Council.

WILLMOTT, A.S. (1994) *Personal communication.*

WILMUT, J. and ROSE, J. (1989) *The Modular TVEI Scheme in Somerset: its concept, delivery and administration.* Report to the Training Agency of the Department of Employment, London.

WILMUT, J. (1996) Personal communication.

WILSON, S.M and GUDMUNDSDOTTIR, S. (1987) What is this a case of? Exploring some conceptual issues in case study research. *Education and Urban Society*, Vol. 20, No. 1, pp. 42-54.

WOLCOTT, H.F. (1990) *Writing up Qualitative Research, Qualitative Research Methods Series 20.* Newbury Park, CA, Sage Publications.

WOLF, A. (1993) *Assessment Issues and Problems in a Criterion-based System.* London, Further Education Unit.

WOLF, T.H. (1972) *Alfred Binet.* Chicago, University of Chicago Press.

WOOD, R. (1986) The agenda for educational measurement. In D. Nuttall. (Ed.) *Assessing Educational Achievemen.,* London, Falmer Press.

WOOD, R. (1982) Aptitude and achievement, *Caribbean Journal of Education*, 9, 79-123.

WOOLNOUGH, B. (1991) *The Making of Engineers and Scientists.* Oxford, University of Oxford Department of Educational Studies.

YATES, A. and PIDGEON, D. (1957) *Admission to Grammar Schools.* London, Newnes.

Appendix 1

Assessment Grids WJEC GCSE 1994 Physics and Biology

PHYSICS

Relationship between Assessment Objectives and Content

Assessment Objectives / Content	3.1 Knowledge/ Recall	3.2 Understanding	3.3 Processes (including experimental skills—see section 6.3)	Mark allocations in complete assessment	
Matter	5%	5%	5%	Not less than 15%	Not less than 60%
Energy	5%	5%	10%	Not less than 15%	
Interactions	5%	10%	10%	Not less than 15%	
Physical quantities	5%	0%	0%	Not less than 5%	
Extension of or addition to the core	5%	10%	20%	One third of total marks	
Mark allocations in complete assessment	Of the order of 45% (at least 20% to understanding)		Not less than 40% (at least 20% to experimental skills)		

The fifth column of the Content section gives a list of examples of devices or situations in which features of the corresponding topic have been usefully applied. The list is not meant to be definitive nor does it claim to include only the best examples.

BIOLOGY

Relationship between assessment objectives and content

Assessment Objectives	Knowledge and Understanding	Skills and processes (including experimental and observational skills.)	Mark allocation in complete assessment
Diversity of organisms.			5 - 10%
Organisation and maintenance of the individual.			40%
Development of organisms and the continuity of life.			25%
Relationships between organisms and with the environment.			25 - 30%
Mark allocation in complete assessment.	Of the order of 45%	Not less than 40% (at least 20% to experimental and observational skills.)	

At least 15% of the marks in any complete examination will be allocated to topics related to the personal, social, economic and technological applications of Biology in modern society (see objectives 2.1 (ii) and 2.5).

Appendix 2

Summary of Mark Weightings allocated to Different Cognitive Demands 1993 - 1995 WJEC GCSE Biology, Chemistry and Physics Examination Papers

		Percentage of the whole paper's marks allocated to a type of cognitive demand		
Examination	**Science Subject**	**Knowledge**	**Comprehension and Application**	**Analysis, Synthesis and Evaluation**
1993	**Biology**	45	37	18
	Chemistry	62	32	6
	Physics	27	70	3
1994	**Biology**	50	33	17
	Chemistry	75	22	3
	Physics	34	53	13
1995 Tier 03	**Biology**	37	56	7
	Chemistry	43	43	14
	Physics	60	32	8
1995 Tier 02	**Biology**	52	45	3
	Chemistry	49	42	9
	Physics	53	43	4

Taken from: *The cognitive skill demands of WJEC GCSE Science Examination Papers,* Benson A., 1995. University of Oxford, Oxford.

Appendix 3

Profile of the Examination Centres Associated with this Study

(Numbers in brackets = No. of students in the centre entered for the WJEC biology, chemistry and physics examinations in this study)

Centre Identities			
1993	1994	1995(option 03)	1995(option 02)
68138 (12)	68138 (11)	68138 (7)	68138 (9)
		68140 (22)	68140 (16)
			68189 (5)
	68220 (23)	68220 (22)	68220 (13)
68221 (18)	68221 (21)	68221 (17)	68221 (4)
68223 (20)	68223 (18)	68223 (9)	68223 (3)
68254 (12)	68254 (10)		
68258 (12)	68258 (16)	68258 (10)	68258 (4)
68286 (5)	68286 (1)		
68303 (4)	68303 (4)		68303 (11)
68304 (5)		68304 (2)	
68306 (40)	68306 (40)	68306 (31)	68306 (25)
68308 (19)	68308 (15)	68308 (7)	68308 (26)
68310 (2)	68310 (4)		
68312 (20)	68312 (34)	68312 (18)	68312 (53)
68317 (9)	68317 (18)	68317 (2)	68317 (21)
68326 (21)	68326 (22)	68326 (9)	68326 (24)
	68329 (1)		68329 (1)
	68332 (23)	68332 (7)	68332 (9)
	68333 (12)	68333 (4)	68333 (19)
	68341 (13)	68341 (4)	68341 (15)
68348 (21)	68348 (23)	68348 (18)	68348 (2)
			68352 (27)
68354 (5)	68354 (9)	68354 (4)	68354 (11)
68358 (12)	68358 (9)	68358 (2)	68358 (10)
68365 (1)			
68369 (12)	68369 (10)	68369 (3)	68369 (13)
68371 (17)	68371 (8)		68371 (25)
68374 (7)	68374 (7)		
		68378 (3)	68378 (5)
	68376 (6)		
	68407 (14)	68407 (4)	68407 (11)
68409 (6)	68409 (5)	68409 (7)	68409 (2)
68411 (5)		68411 (2)	68411 (1)
68507 (8)	68507 (7)	68507 (9)	68507 (9)
68510 (5)	68510 (9)		
68516 (31)	68516 (17)		
68520 (19)	68520 (31)	68520 (17)	68520 (17)
		68521 (7)	68521 (8)
			68524 (1)
		68535 (6)	68535 (17)
68538 (10)	68538 (12)		
68544 (2)			
68545 (16)	68545 (9)	68545 (5)	68545 (14)

Continued

1993	1994	1995(option 03)	1995(option 02)
	68550 (6)		68550 (9)
68558 (17)	68558 (13)	68558 (10)	68558 (11)
68560 (11)	68560 (12)	68560 (8)	68560 (7)
	68564 (1)		
	68566 (11)		
		68567 (6)	68567 (10)
		68573 (11)	68573 (6)
	68575 (6)	68575 (9)	68575 (10)
68570 (1)			
68589 (13)	68589 (9)	68589 (8)	68589 (9)
68609 (1)			68609 (6)
		68635 (6)	68635 (17)
68643 (10)	68643 (11)	68643 (11)	68643 (6)
		68656 (9)	68656 (5)
	68677 (12)	68677 (3)	68677 (10)
		68685 (9)	68685 (5)
	68719 (8)	68719 (9)	68719 (4)
		68722 (1)	68722 (19)
68739 (12)	68739 (11)	68739 (1)	68739 (6)
68740 (6)	68740 (1)		
68741 (3)	68741 (5)		
68743 (1)	68743 (1)		
68745 (2)			68745 (1)
		68751 (5)	68751 (4)
68761 (12)			
68765 (22)	68765 (18)		
	68767 (1)		
68775 (14)	68775 (14)		
68785 (8)	68785 (7)	68785 (4)	68785 (2)
68816 (4)	68816 (4)		
68818 (4)	68818 (16)	68818 (4)	68818 (15)
	68825 (11)		
	68827 (21)		
68835 (1)	68835 (6)		
68846 (19)	68846 (8)	68846 (4)	68846 (13)
	68854 (9)		
68855 (24)	68855 (39)		
68866 (8)	68866 (8)	68866 (11)	68866 (4)
	68868 (12)		
68870 (49)	68870 (43)		
68871 (13)	68871 (5)		
	74000 (1)		
Centres (N=53)	Centres (N=64)	Centres (N=48)	Centres (N=55)

(For 1995, Options 03 and 02 have 46 Centres in common, and between them offer a total of 56 different centres)

Appendix 4

STATUS AND TYPE OF CENTRE

	1993 No. of Centres	1993 % of Population's Students	1994 No. of Centres	1994 % of Population's Students	1995(03) No. of Centres	1993(03) % of Population's Students	1993(02) No. of Centres	1993(02) % of Population's Students
Secondary (Comp.) and Middle	37	85.6	52	91.0	39	86.3	43	92.6
Secondary (Independent)	12	12.5	10	8.0	7	11.4	8	4.8
Further Education Establishment	1	0.2	0	0.0	0	0.0	2	0.9
Sixth Form College	1	1.2	1	0.9	1	2.3	1	1.5
Tertiary College	2	0.5	0	0.0	0	0.0	1	0.2
Unclasified	0	0.0	1	0.1	0	0.0	0	0.0

LOCUS OF CONTROL

	1993 No. of Centres	1993 % of Population's Students	1994 No. of Centres	1994 % of Population's Students	1995(03) No. of Centres	1993(03) % of Population's Students	1993(02) No. of Centres	1993(02) % of Population's Students
Maintained(LEA/Ed/Authority	35	84.5	49	88.6	39	86.3	44	9286
Independent	12	12.5	10	8.0	7	11.4	8	4.8
Aided(Voluntary aided and aided)	1	0.8	2	1.9	0	0.0	0	0.0
Grant Maintained Trust	1	0.3	1	0.5	0	0.0	0	0.0
SCEA (Forces Centres overseas	1	1.3	1	0.9	1	2.3	1	1.5
Unclasified	3	0.6	1	0.1	0	0.0	2	0.9

AGE RANGE OF STUDENTS

	1993 No. of Centres	**1993** % of Population's Students	**1994** No. of Centres	**1994** % of Population's Students	**1995(03)** No. of Centres	**1993(03)** % of Population's Students	**1993(02)** No. of Centres	**1993(02)** % of Population's Students
11/12	12	18.7	15	15.5	6	5.9	8	17.2
11/12 – 18/19	29	70.3	39	76.4	34	82.4	37	75.6
13/14 – 16	1	1.3	1	1.0	1	1.1	1	0.3
13/14 – 18/19	1	1.3	2	2.5	2	3.4	2	3.9
16 – 18/19	0	0.0	1	1.0	1	2.3	1	0.6
16 - Adult	3	0.6	0	0.0	0	0.0	3	1.2
5 – 16/18	7	7.8	5	3.5	3	4.9	3	1.2
Unclassified	0	0.0	1	0.1	0	0.0	0	0.0

INTAKE IN TERMS OF SEX

	1993 No. of Centres	**1993 % of** Population's Students	**1994** No. of Centres	**1994 % of** Population's Students	**1995(03)** No. of Centres	**1993(03) % of** Population's Students	**1993(02)** No. of Centres	**1993(02)** % of Population's Students
Boys only	1	1.7	1	1.5	1	2.1	1	1.2
Girls only	3	2.4	2	1.0	1	2.3	1	1.5
Boys and Girls	49	95.9	61	97.4	45	95.6	53	97.3

Appendix 5 Descriptive Statistics for WJEC GCSE Biology, Chemistry and Physics

1993 Number of students in each subject, N = 631

Grade	Biology			Chemistry			Physics		
	Freq	%	Cum%	Freq	%	Cum%	Freq	%	Cum%
A	171	27.1	27.1	309	49.0	49.0	175	27.7	27.7
B	199	31.5	58.6	208	33.0	81.9	173	27.4	55.2
C	147	23.3	81.9	91	14.4	96.4	130	20.6	75.8
D	100	15.8	97.8	20	3.2	99.5	129	20.4	96.2
E	3	0.5	98.3	2	0.3	99.8	19	3.0	99.2
F	0	0.0	98.3	1	0.2	100.0	0	0.0	99.2
G	0	0.0	98.3	0	0.0	100.0	0	0.0	99.2
U	11	1.7	100.0	0	0.0	100.0	5	0.8	100.0
Mean			2.40			1.73			2.48
Std. Dev.			1.29			0.86			1.28

1994 Number of students in each subject, N = 792

Grade	Biology			Chemistry			Physics		
	Freq	%	Cum%	Freq	%	Cum%	Freq	%	Cum%
A*	141	17.8	17.8	153	19.3	19.3	135	17.0	17.0
A	247	31.2	49.0	281	35.5	54.8	278	35.1	52.1
B	163	20.6	69.6	207	26.1	80.9	238	30.1	82.2
C	107	13.5	83.1	117	14.8	95.7	104	13.1	95.3
D	131	16.5	99.6	32	4.0	99.7	35	4.4	99.7
E	2	0.3	99.9	0	0.0	99.7	2	0.3	100.0
F	0	0.0	99.9	2	0.3	100.0	0	0.0	100.0
G	0	0.0	99.9	0	0.0	100.0	0	0.0	100.0
U	1	0.1	100.0	0	0.0	100.0	0	0.0	100.0
Mean			1.81			1.49			1.54
Std. Dev.			1.36			1.11			1.07

1995(03) Number of students in each subject, N = 387

Grade	Biology			Chemistry			Physics		
	Freq	%	Cum%	Freq	%	Cum%	Freq	%	Cum%
A*	132	34.1	34.1	103	26.6	26.6	182	47.0	47.0
A	173	44.7	78.8	167	43.2	69.8	150	38.8	85.8
B	74	19.1	97.9	98	25.3	95.1	52	13.4	99.2
C	8	2.1	100.0	19	4.9	100.0	3	0.8	100.0
Mean			0.89			1.09			0.68
Std. Dev.			0.78			0.84			0.73

1995 (02) Number of students in each subject = 610

Grade	Biology			Chemistry			Physics		
	Freq	%	Cum%	Freq	%	Cum%	Freq	%	Cum%
A				5•	0.8	0.8	17•	2.8	2.8
B	217	35.6	35.6	263	43.1	43.9	409	67.0	70.1
C	290	47.5	83.1	247	40.5	84.4	110	18.0	88.2
D	97	15.9	99.0	78	12.8	97.2	60	9.8	98.0
E	3	0.5	99.5	15	2.5	99.7	10	1.6	99.7
F	0	0.0	99.5	1•	0.2	99.8	1•	0.2	99.8
G	0	0.0	99.5	0	0.0	99.8	0	0.0	99.8
U	3	0.5	100.0	1	0.2	100.0	1	0.2	100.0
Mean			2.84			2.74			2.42
Std. Dev.			0.80			0.83			0.82

•indicates the award of an exceptional grade to allow for mistakes in allocation of students to tiers

Appendix 6 Descriptive Statistics for SEG GCSE Biology, Chemistry and Physics

1994

Grade	Biology			Chemistry			Physics		
	Freq	%	Cum%	Freq	%	Cum%	Freq	%	Cum%
A*	109	10.9	10.9	133	13.3	13.3	114	11.4	11.4
A	247	24.7	35.6	256	25.6	38.9	187	18.7	30.1
B	301	30.1	65.6	254	25.4	64.2	246	24.6	54.6
C	225	22.5	88.1	185	18.5	82.7	256	25.6	80.2
D	77	7.7	95.8	119	11.9	94.6	127	12.7	92.9
E	25	2.5	98.3	41	4.1	98.7	51	5.1	98.0
F	8	0.8	99.1	8	0.8	99.5	14	1.4	99.4
G	4	0.4	99.5	4	0.4	99.9	4	0.4	99.8
U	3	0.3	99.8	0	0.0	99.9	1	0.1	99.9
NRG	2	0.2	100.0	1	0.1	100.0	1	0.1	100.0
Mean	2.07			2.08			2.34		
SD	1.37			1.45			1.47		

Number of students in each subject = 1001
NRG denotes no recorded grade

1995

Grade	Biology			Chemistry			Physics		
	Freq	%	Cum%	Freq	%	Cum%	Freq	%	Cum%
A*	376	21.4	21.4	489	27.8	27.8	589	33.4	33.5
A	832	47.2	68.7	631	35.8	63.7	804	45.8	79.4
B	464	26.3	95.1	420	23.9	87.5	298	16.9	96.3
C	81	4.6	99.7	167	9.5	97.0	62	3.5	99.8
U	6	0.3	100.0	52	3.0	100.0	3	0.2	100.0
Mean	1.17			1.36			0.92		
SD	0.90			1.49			0.85		

Number of students in each subject = 1759

Appendix 7

INTERVIEW SCHEDULE / AIDE MEMOIRE

School/Date .....

Teacher .....

Teaching Speciality/ Years Teaching Experience/ Degree Speciality
ONLY ENTER AT THE END OF THE INTERVIEW
...../...../.....

Interviewer remember! - scan the expected issues that have been covered before moving onto next question so as to continually monitor progress and course of the interview. Expected issues are those that I think might emerge in the interviews – those shown in bold are the main issue/s that I wish to cover.

CHECK TEACHER HAS MARK BOOK

1 **I'd like us to begin by asking you to tell me about the Year 9 groups and classes you teach. How many classes do you teach / do you teach them all three sciences / is your teaching of them shared with other science staff / are they mixed ability or set / did you teach them in Years 7 and 8 /**
How did pupils get into their Year 9 science teaching groups or classes?

Expected issues

-Extent of teacher's involvement in Years 7, 8, 9 science teaching (match later with total number of groups/classes in each KS3 year and with teacher's own KS4 involvement).

-Teacher's experience of teaching biology/chemistry/physics (match later with their own specialism and what they teach at KS4).

-Principles/Processes used for allocating pupils to teaching groups (is a pupil's allocated KS4 science actually 'set up' from their placement in year 7/8/9?)

2 **Tell me about the KS4 groups/classes that you teach. How many classes do you teach / do you teach them all three sciences / is your teaching of them shared with other science staff / are they mixed ability or set / did you teach them in Year 9/ <u>How did pupils get into their KS4 science teaching groups or classes?</u> <u>What exams do they do?</u> <u>Do pupils move between these groups/classes in KS4 – when, why and who decides?</u>**

Expected issues

-Extent of teacher's involvement in Years 10 and 11 science teaching (match later with total number of groups/classes in each KS4 year)

-Teacher's experience of teaching biology/chemistry/physics (match later with their own specialism)

<u>-Principles/Methods used for allocating pupils to teaching groups.</u>

**Are CAT or SAT scores influential in pupils' KS4 group/class science allocations – what else is influential in this respect?*

**Pupils' science or separate sciences (biology/chemistry/physics) group and class placements in KS4. Whose decision / what policy? What say does the class teacher have in pupils' group/class placement?*

**Is a pupil's allocated KS4 science group/class actually 'set up' from their placement in Year 7/8/9?*

**Movement of pupils between groups/classes – in theory or a reality – when, why and who decides?*

**Have there been or are there going to be changes in the above practice / policy and if so, why and when?*

CHECK TAPE AND TIME!

3 HAVE REASONS FOR USED / ALLOCATED
GCSE EXAMINATION GROUP
SYLLABUS
TIER
EMERGED YET? NOTE THAT PATTERN OF USE OVER 1990S MAY NOT HAVE EMERGED YET. If not then ask the following questions now:

Tell me about the particular GCSE examination group that your pupils are entered with? Have you used the same GCSE examination group and the same syllabuses throughout the 1990s?

Expected issues

*Reasons for WJEC/ other examination group

*If there have been changes WHAT / WHO has prompted them and what is the TIME FRAME in this respect?

CHECK TAPE AND TIME!

4 HAVE REASONS FOR USED / ALLOCATED TIERS EMERGED YET?

IF NOT, turn the teacher's attention to this issue by using the following task.

We now need to refer to your class lists / mark book. I'm going to randomly pick a boy and a girl in one of your KS4 teaching groups and ask you to tell me how each pupil came to sit the science exam that they did this summer or are allocated to sit next summer. NOTE: the number of boys/girls will vary according to how many groups the teacher takes/ the richness of what has already emerged/the willingness of the interviewee to continue with the interview and engage in this process.

Expected issues

*Policy and practice issues should emerge again with respect to pupils' allocated examination papers. Hopefully the reality of what actually happens in the school should emerge.

*Tiering choices and reasons for them. Whose choice? **Parental pressure mentioned ?**– if not, **ask them to talk about any factors other than those within school which play a part in deciding which syllabus/tier a pupil is entered for.**

***Movement between groups** might emerge here if it has not done so already. **If it does not, ask directly if movements do occur between teaching groups and if so to what extent and when / upon what do the movements depend / who decides that it will happen?**

*Gender-related issues should emerge here if they have not done so already. If not, leave until the outcomes of Question 5.

*Science subject difficulty might emerge here if not done so already. In any event continue with Question 5 and use anything that has been raised as a way into this question.

CHECK TAPE AND TIME!

5 Some people believe that the separate science GCSEs are not equally difficult for pupils. What do you think?

Expected issues

RESPONSE: YES, they do differ in difficulty

Issues to explore if they do not emerge, or emerge only superficially:

**what is the perception and is it based upon anecdote or hard evidence (explore any evidence offered by teacher e.g. this year's exam papers and related factors)*

**persistency of difference / pattern over the 1990s*

**impact of NC in early to mid-1990s/ revision of NC post Dearing*

**coursework / paper construction factors as factors influencing 'difficulty'*

****gendered effects** – be particularly careful that I do not use words/phrasing that pre-empts what teachers might say here.*

**introduction of A* effect on different separate sciences at GCSE*

**factors related to tiering and the process of pupil allocation*

RESPONSE: NO, they do not differ in difficulty

I'm not expecting this response but if given, I'll continue with the following to tease out teachers' perceptions.

How would you respond to people that claim:

** physics is harder because of the greater proportion of marks allocated to calculation work or that the calculation work is harder than in the other sciences.*

**chemistry is harder because it requires pupils to think more about abstract things.*

**biology is harder because there's more to remember.*